Mechanisms of Ensemble Face Processing: Extraction of Summary, but not Single, Identity Shows Sensitivity to Non-Frontal Global Viewpoints

by

Marco Agazio Sama

A thesis submitted in conformity with the requirements for the degree of Master of Arts
Department of Psychology
University of Toronto

© Copyright by Marco A. Sama 2017
Mechanisms of Ensemble Face Processing: Extraction of Summary, but not Single, Identity Shows Sensitivity to Non-Frontal Global Viewpoints

Marco A. Sama

Master of Arts

Department of Psychology
University of Toronto

2017

Abstract

The recognition and perception of faces – both alone and in groups – is an important biological function, enabling humans to effectively engage with one another socially. This thesis describes viewpoint and identity interactions in processing face ensembles, helping vision scientists to understand the cognitive mechanisms taking place. I manipulated viewpoint or identity across two experiments. The first experiment had participants report an average or single viewpoint or identity, and the second experiment had participants report an average identity or a single identity across a wide range of viewpoints. I find that efficiency in average – but not single – identity extraction decreases as viewpoints are progressively non-frontal. Viewers also extract summary viewpoint better than a single viewpoint, and this ability also decreases as global viewpoints are non-frontal. The implications in the face ensemble processing system, general ensemble processing models, and future research are discussed below.
Acknowledgments

I would like to acknowledge the role of Dr. Adrian Nestor and Dr. Jonathan S. Cant in not only supervising this research project, but also patiently supporting my transition into graduate studies. Of my now five years at the University of Toronto Scarborough, my Master’s year was the most enjoyable in part because of their welcoming lab environments. I look forward to the next four (or more) years during supervision of my doctoral degree. I would also like to thank various lab members (in no particular order) who helped me familiarize myself with the role of a graduate student researcher in the lab: Mr. Sol Sun, Mr. Matt Lowe, Mr. Ashutosh Patel, Ms. Natasha Drobotenko, and Dr. Dan Nemrodov, among various other individuals.

I also owe thanks to the research assistants who helped me collect, input, and organize the data: Mr. Sajeevan Sujanthan, Ms. Janannii Selvanathan, Mr. Dilakshan Srikanthan, Ms. Adile Nexha, Ms. Megan Saliani, Ms. Annabel Fan, and Mr. Zachary Grey.

I would also like to acknowledge those who supported me during my undergraduate years, without whom I probably would not have achieved the academic success required to enter graduate school. Specifically, Dr. Cindy Bongard for being a mentor for most of my undergraduate degree, and Dr. Konstantine Zakzanis for giving me my first exposure to the research environment which set in stone my interest to pursue graduate studies.

Finally, I would like to acknowledge Mrs. Suzie Li, my wife of six years, who has supported my journey through academia. Without your help, I can say with high certainty that I may never have had the chance to pursue academia, or even attend university.
Table of Contents

Acknowledgments ........................................................................................................ iii

Table of Contents ....................................................................................................... iv

List of Tables .............................................................................................................. v

List of Figures ........................................................................................................... vi

Chapter 1: Introduction and Background .................................................................... 1

1.1 Preamble .............................................................................................................. 1

1.2 Single Face Processing ....................................................................................... 2

1.3 Why Study Ensembles? ..................................................................................... 5

1.4 Purpose of this Research .................................................................................. 8

Chapter 2: Methods and Results of Experiments 1 and 2 ......................................... 12

2.1 General Methods .............................................................................................. 12

2.2 Experiment 1: Comparison of Individual and Summary Extraction of Viewpoint and Identity from Face Ensembles ................................................................. 16

2.3 Experiment 2a: Viewpoint Sensitivity in the Extraction of Summary Identity from Face Ensembles ................................................................................................. 23

2.4 Experiment 2b: Viewpoint Invariance in the Extraction of Single Face Identity from Face Ensembles .................................................................

Chapter 3: General Discussion and Future Directions ............................................. 31

3.1 Highlights of General Findings ......................................................................... 31

3.2 Implications for Future Research ..................................................................... 32

3.3 Limitations ......................................................................................................... 35

3.4 Conclusion ......................................................................................................... 37

References ................................................................................................................. 49
List of Tables

TABLE 1 ......................................................................................................................... 39
List of Figures

Figure 1. ................................................................................................................. 40
Figure 2. ................................................................................................................. 41
Figure 3. ................................................................................................................. 42
Figure 4. ................................................................................................................. 43
Figure 5. ................................................................................................................. 44
Figure 6. ................................................................................................................. 45
Figure 7. ................................................................................................................. 46
Figure 8. ................................................................................................................. 47
Figure 9. ................................................................................................................. 48
Chapter 1
Introduction and Background

1.1 Preamble

Imagine a subway train, lecture theatre, or any other commonly crowded location. From the people around you, you can gather important nonverbal social information. Directional cues and identity features can be extracted from the general movement of the crowd. When disembarking the crowded train, these cues can hint towards an exit or transfer. If you are standing in a populated lecture theatre, you can grasp how engaged the students are simply by their collective gaze or expression. In another example, if you were watching a large social gathering, you could ascertain whether the gathering was celebratory or mournful based on the collective expressions of those participating and without any audio cues. This is all seemingly done effortlessly and automatically. Meanwhile, the human brain is engaging in complex neural computations and many of the mechanisms underlying these cognitive processes have yet to be fully understood by vision scientists.

The few examples stated above illustrates how perceptual processing of faces serves an extremely important biological function for humans. The fluency of attaining relevant information from faces allows us as a species to engage in effective communication with one another. The recognition of one another and collective group dynamics seems to be a functional requirement for social survival, and is even found in nonhuman eusocial species demonstrating a potential evolutionary adaptation. For example Polistes fuscatus, a species of social wasps, show better recognition for faces of other P. fuscatus members than different visual stimuli (Tibbetts, 2002; Sheehan & Tibbetts, 2011), and this face-specialization has been shown to be genetically driven based on genomic comparisons with a cousin species P. metricus which does not show face-specialization (Berens, Tibbetts, & Toth, 2017). Closer to humans, the primate Macaca mulatta has been shown to discriminate faces of both humans and fellow primates, with increased fixation to novel primate faces than human ones (Pascalis & Bachevalier, 1998). For social species that are less visually-inclined, selective individuation and recognition of group-membership can occur through other sensory systems, such as olfaction in the social ant Camponotus laevigatus (Sharma et al., 2015).
The ability of a species to recognize individual members and categorize groups is entwined its ability to be social. One could even parallel it with a “which came first, the chicken or the egg” analogy. Did increase in sociability drive the evolution of specialization in member recognition, or did increase in member recognition drive the evolution of enhanced sociability? This question, albeit philosophically and scientifically interesting, is beyond the main focus of this thesis. Here, I focus on the cognitive mechanisms underlying summary face extraction from crowds of faces – or ensembles. This is part of the driving process in the ability for humans – and other social animals – to efficiently extract information from other members.

1.2 Single Face Processing

Faces are a special case of visual stimuli. Neural and cognitive processing of human faces happens holistically. This means the stimulus is processed as a whole rather than built up from recognizing individual components, and is evident by the composite face illusion (Rossion, 2013; Taubert & Alais, 2009; Jacques & Rossion, 2009; Schiltz, Dricot, Goebel, & Rossion, 2010; Laguesse & Rossion, 2013; Richler & Gauthier, 2014), and the face inversion effect (Maurer, Le Grand, & Mondloch, 2002; Sekuler, Gaspar, Gold, & Bennett, 2004). Where specialization in visual recognition of complex objects like cars or birds requires training, we seem to possess an innate specialization for recognizing faces (Simion & Di Giorgio, 2015). In humans, functional imaging research has localized what appears to be a specialized face-processing region in the ventral temporal lobe dubbed the fusiform face area, or FFA (Kanwisher, McDermott, & Chun, 1997; McCarthy, Puce, Gore, & Allison, 1997; Kanwisher & Yovel, 2006). Recent findings also point to another face-specific area in the occipital lobe: the occipital face area or OFA (Pitcher, Walsh, & Duchaine, 2011; Kietzmann et al., 2015). From here, other facial-recognition cortical structures have been identified, including the anterior face patch in the anterior temporal lobe (Rajimehr, Young, & Tootell, 2009; Von Der Heide, Skipper, & Olson, 2013), and the superior temporal sulcus (Nguyen & Cunnington, 2014). Early research also showed widespread extrastriate cortical activation to face stimuli (Haxby et al., 2001), claiming a distributed representation for processing faces. Macaque monkeys, evolutionarily similar to humans, also show discriminability to across human face exemplars from patches of face cells (Chang & Tsao, 2017). These findings merely highlight the vast literature on the cognitive and neural mechanisms of human face processing (for a review see Anzellotti & Caramazza, 2014). Where early researchers tended to argue for either strong localized activation (e.g. Kanwisher et al.,
1997) or strong distributed representation (e.g. Haxby et al., 2001), the current understanding is more hybrid, where there are local centers for face processing but also extrastriate involvement organized as a face-processing network (Ishai, 2008; Zhen, Fang, & Liu, 2013; Nestor, Plaut, & Behrmann, 2011; Anzellotti, Fairhall, & Caramazza, 2014) with vast interconnectivity (Fairhall & Ishai, 2007).

One of the largest ongoing questions in the literature pertains to whether face processing expertise is innate or learned. Proponents of a learned-hypothesis may argue that we are constantly exposed to human faces from birth, making us develop a stronger expertise for faces than other visual stimuli. Probably the most famous phenomenon surrounding this hypothesis is the acquired expertise of the novel object Greebles, which when viewed activates FFA (Gauthier, Tarr, Anderson, Skudlarski, & Gore, 1999) and evokes the face-specific N170 event-related potential (ERP) response (Rossion, Gauthier, & Goffaux, 2002). FFA activation also occurs when car and bird experts passively view the stimulus they have expertise with (Gauthier, Skudlarski, Gore, & Anderson, 2000). An opponent of the learned-hypothesis may suggest that FFA activation to birds and cars is due to the face and bodies of the former and the grill-headlight arrangement resembling a face in the latter. However, FFA activation has also been shown when radiologists view X-ray scans (Bilalić, Grottenthaler, Nägele, & Lindig, 2016), images which are non-face like. Even in this case, however, we cannot rule out activation of FFA because of the subsidiary information that X-ray scans carry. Radiologists know they are looking at bodies, which could be driving FFA activation (Peelen & Downing, 2005). Additionally, FFA activation to acquired stimulus expertise does not rule out the innate role of the FFA in recognizing faces. Instead, FFA – because of its established expertise at analyzing faces which have local variations in homogenous global arrangements – can be recruited for other stimuli with similar global-local arrangements. For example, all cars will have two headlights at each side and usually a central-grill, but the local shapes of each may differ – just like faces. Even biological radiographs share this feature. The anatomical arrangement in each image will be the same, but the local sizes and shapes will differ and it is the radiologist’s job to discern if the local shape is varying enough to suggest an abnormality.

There are also ongoing debates on the nature of Greebles, with some saying they activate FFA because they are face-like (Brants, Wagemans, & Op de Beeck, 2011). However, others have argued from neuropsychological research that Greebles are not face like (Gauthier, Behrmann, &
Tarr, 2004). Here, a visual agnosic patient was unable to distinguish Greebles despite having intact face processing, and two prosopagnosic patients displayed the opposite results – preserved ability to differentiate Greebles despite loss of face identity processing. The debates surrounding Greebles highlight an important gap in understanding many of the mechanisms taking place in FFA and other cortical structures to both learned and innate processing expertise. While outside the main focus of this thesis, they provide an important context from which researchers can understand the mechanisms of human face processing.

Proponents of the innate-hypothesis argue that human visual cortex develops face recognition automatically, likely during prenatal development. Newborn infants already show both discriminability for human faces and the ability to distinguish between face-like stimuli and other stimuli, and even show the face-specific N170 ERP signal (de Haan, Pascalis, & Johnson, 2002). However, infants show invariance to inverted faces, demonstrating that the face-processing network is not fully developed postnatally (de Haan et al.). Evidence for an innate-mechanism is also suggested by the existence of congenital prosopagnosia – the inability to recognize faces from birth which does not seem to ameliorate over time (for a review see Behrmann & Avidan, 2005). It is likely that humans are born with existing face processing abilities, but this system is not fully developed and requires fine-tuning during postnatal development (Simion & Di Giorgio, 2015).

The learned versus innate hypothesis of facial recognition continues to be an ongoing debate, beyond the scope of what is reviewed here. We cannot discount FFA activation to non-face stimuli, but this does not necessarily rule out the innate processing expertise to human faces. As newer information emerges, it can hopefully be of use by vision scientists to refine models of human face processing. An innate route may suggest genetic mechanisms leading to cortical regions specialized for processing faces (e.g. Wilmer et al., 2010), while a learned route suggests plasticity in visual cortex to process faces and other objects of expertise (e.g. Tanaka & Pierce, 2009). Of course, the existence of one route does not negate the other. In the context of other findings, human face processing is important for the successful functioning of social species which explains why expertise for other anatomical landmarks – such as hands – fail to develop in visual cortex despite similar exposure to both hands and faces (McKone, Kanwisher, & Duchaine, 2007).
1.3 Why Study Ensembles?

The majority of face processing research focuses on single-face stimuli, usually presented in foveal vision. This is contrasted with exposure to faces in the real world. While it is not uncommon to view single-faces foveally – such as a conversation with a spouse or colleague – in our day-to-day activities we often encounter faces as part of a crowd and research has only recently begun to investigate how processing of faces differs when they are part of an ensemble. Even in these situations, humans show remarkable capability to extract relevant information. Within 50 ms, the human visual system can discern whether two serially presented faces are the same or not (Richler, Mach, Gauthier, & Palmeri, 2009), and recent pattern analysis techniques have found identity-specific ERP signatures as early as 70 ms (Nemrodov, Niemeier, Mok, & Nestor, 2016), demonstrating fast identity individuation. However, in many daily situations we encounter more than two simultaneous faces. Having a large number of simultaneous stimuli poses an interesting problem to the limitations of the visual system (Cowan, 2010). Our visual working memory (VWM) appears limited in its ability to hold information about individual exemplar stimuli. As the number of objects in the visual field increases, the ability to capacitate each one in detail diminishes. Generally, it has been thought that the visual system can hold three-to-four items at a time (Luck & Vogel, 1997; Raffone & Wolters, 2001). However, debates are currently ongoing regarding the actual capacity as well as whether VWM contains discrete slots or continuous resources for allocating information about any exemplar (Luck & Vogel, 2013). Ensemble processing circumvents this limitation by extracting a statistical summary of the object set. In one of the earliest ensemble studies, Ariely (2001) demonstrated that observers can extract a fairly accurate average size representation of a set of circles, but were poor at reporting the size of any one particular circle. This summary extraction partially explains the phenomenon of having a rich perceptual world, despite the inability to accurately process all of it (Cohen, Dennett, & Kanwisher, 2016; Yamanashi Leib, Kosovicheva, & Whitney, 2016). The ability to extract a summary of the set with more accuracy than any individual item in large sets is a key feature of ensemble processing (for a general review of ensemble processing, see Alvarez, 2011). In addition to summary representation, research has also found sensitivity to variance within feature dimensions in ensembles (Haberman, Lee, & Whitney, 2015; Solomon, 2010).
Not only does ensemble processing help us with overcoming VWM limitations, but it also helps us quickly categorize sets of visual stimuli based on shared, combinable, features (Utochkin, 2015). Ensemble processing benefits humans by allowing the visual system to quickly summarize large quantities of information, giving us the ability to quickly visualize trends (Szafir, Haroz, Gleicher, & Franconeri, 2016). Without ensemble processing, the very act of scientific investigation may be near unattainable as this process seems to drive our ability to interpret arrays of data in figure format (Szafir et al.). Nevertheless, ensemble processing is not free from its own set of limitations, much like VWM capacity for single items. For example, it appears that we can only store a couple ensemble features simultaneously (see Utochkin).

Behavioural research has demonstrated distinct cognitive mechanisms pertaining to processing single objects versus object ensembles (Cant, Sun, & Xu, 2015). Here, response time (RT) to changes in one attended feature for single objects was not impacted by changes in the other feature. For example, observers could ignore change in shape if they were attending to change in texture, and vice-versa. However, for object ensembles, observers could not ignore a change in an unattended feature, resulting in a longer RT when an unattended feature changed in addition to the attended feature. While to my awareness no functional imaging studies have directly compared single and object ensembles, two studies showed shape and texture for single objects activates in the lateral occipital complex (LOC) and collateral sulcus (CoS), respectively (Cant & Goodale, 2007; Cant, Arnott, & Goodale, 2009); on the other hand shape and texture for object ensembles shows activation in scene-selective parahippocampal place area (Cant & Xu, 2017). However, caution should be warranted in comparing those findings as they contained different stimuli and paradigms. Given distinctions found by direct behavioural comparisons (Cant et al., 2015), it is likely that a direct neural comparison would be consistent with the above two findings, but future research is still needed to demonstrate this empirically.

Other research has shown a distinction between ensemble processes for high level versus low level feature domains, such as face identity versus gabor orientation (Haberman, Brady, & Alvarez, 2015). In this elegant behavioural experiment, performance was correlated both between and within category domains. Here the authors report that performance was more correlated within domains than between them. From this, they argued for a model of ensemble processing which has independent systems for both high level and low level stimuli (see Figure 8 in Haberman, Brady, & Alvarez, 2015).
Ensemble processing has recently been applied to faces, with two early studies demonstrating efficient extraction of summary identity (Haberman & Whitney, 2007; de Fockert & Wolfenstein, 2009). Developmentally, young children already show the ability to extract a statistical summary of faces, and the efficiency increases with age (Rhodes et al., 2017). This seems to suggest that, like single-face processing mechanisms, ensemble processing systems develop at a young age and may already be present in infancy (Zosh, Halberda, & Feigenson, 2011). Children with autism, who show diminished social capabilities, extract average identity poorer than neurotypical children (Rhodes, Neumann, Ewing, & Palermo, 2015). The ability to extract summary information from crowds (where one would experience face ensembles ecologically) provides us with an important tool for socializing and interacting with others. It is important to know whether a crowd of people is upset with you or happy with you, especially in social roles involving interactions with large groups of people simultaneously, such as a politician, teacher, or stage performer. Partaking in a social hierarchy seems to necessitate efficient and relevant extraction of summary social information from groups. Haberman and Whitney (2010) demonstrated this with the finding that observers can filter out deviant emotions when extracting average crowd expression. In this study, observers were presented with 12 faces of varying emotions, with two of those faces having expressions which were highly deviant from the other ten. After the display, observers cycled through a series of emotional expressions on a single probe face to report the average emotion. Rather than reporting an expression which equally averaged all 12 emotional faces, their reports tended to filter out the two outliers. From this, the authors reasoned that obtaining a true global emotion may not be needed when observers sample the expression of a crowd. This also emphasizes Utochkin’s (2015) argument that when rapidly categorizing a set of features, highly deviant ones may be less important to the perceptual summary. A trend-like approach seems to be more socially important. Imagine the politician at a rally: what would be important to them is that the vast majority of the audience is content with their platform. The few naysayers would be filtered out. The same can be said for the university lecturer. As long as most of a large class appears to be engaged in your teaching, you can discount the few disinterested students. This summary extraction also seems to occur automatically (Haberman & Whitney, 2007).

Gaze direction and sex, two additional socially relevant features, can also be extracted from face ensembles (Sweeny & Whitney, 2014; Haberman & Whitney, 2007). Extraction of summary
features does not seem to rely on foveal fixation either. In fact, Wolfe, Kosovicheva, Yamanashi Leib, Wood, and Whitney (2015) found adequate extraction of average expression from faces displayed 200 to 300 pixels (6.75° to 10.5° visual angle) from central fixation.

Overall, it seems that we can extract a summary of almost any face-related feature. Despite understanding what the ensemble processing system can do, it is not well understood how extraction occurs (Alvarez, 2011) including the cognitive mechanism that take place, particularly in ensemble face processing. For example, it is not well established whether ensemble face processing uses the same cognitive mechanisms and cortical networks as single-face processing or completely distinct cortical networks. First, based on behavioural findings it may be that single objects and object ensembles may utilize different networks (Cant et al., 2015). Second, a study of four prosopagnosic subjects showed adequate ensemble processing of faces – comparable to healthy controls (Yamanashi Leib et al., 2012). From this, even if a face-specific ensemble mechanism does exist, it may likely be dissociable from single-face perceptual systems. Furthermore, whether object ensembles and face ensembles are processed by a shared or distinct ensemble processing system is an additional interesting question. Discerning the functional cognitive mechanisms of face ensemble processing is thus an important first step to later forming testable models when it comes to studying anomalies like prosopagnosia or autism, or understanding how ensemble face processing differs from single face processing and non-face object-ensemble processing. The research conducted in this thesis will hopefully help researchers understand the basic cognitive mechanisms underlying ensemble face processing by comparing how low- and high-level face features interact with one another.

1.4 Purpose of this Research

The research in this thesis sheds light on the basic underlying cognitive mechanisms of face ensemble processing, some of which have already been investigated. Discussed above, Haberman, Brady, and Alvarez (2015) reported dissociated mechanisms when processing high level versus low level visual stimuli. One immediate question is whether this independent system refers to high and low level visual stimuli or face and non-face stimuli when processing ensembles. Here, I will expand on this by using high and low level visual features that are both components of faces, specifically identity (high level) and viewpoint (low level), both of which are important basic features of face processing. Some investigation of interactions between the
processing of viewpoint and identity for face ensembles has already been conducted. Yamanashi Leib, Liu, Qiu, Robertson, and Whitney (2014) looked at average identity when faces are presented at various viewpoints. In this study, 144 continuous identities were generated by linearly morphing three base identities. There were 47 morphs between Identities 1 and 2, 47 morphs between Identities 2 and 3, and 47 morphs between Identities 3 and 1. Each of the 144 morphs ranged across five viewpoints: left profile, 22.5° to the left, frontal (facing participants), 22.5° to the right, or right profile. During the experiment, six faces were presented three times serially to participants (total of 18 per trial) and varied in identity and viewpoint. Identities ranged from +25 morphs above a preset mean to -25 morphs below the preset mean, varying in 10 morph units. The actual mean identity was not shown. Participants reported the perceived mean identity by cycling through all the identity morphs in a single probe face. The authors found that participants were able to extract an adequate summary identity across the varying viewpoints. From this, they concluded that average identity extraction is invariant to changes in viewpoint.

Face ensemble encoding mechanisms likely involve some probabilistic processing, making inferences on the average identity to overcome perceptual ambiguity of non-frontal viewpoints (Yamanashi Leib et al., 2014). This is consistent with the recent interest in thinking of the visual system as using Bayesian-like computations when analyzing the visual world, making it appear quite efficient at discerning our everyday surroundings (Sterzer & Kleinschmidt, 2007; Purves, Monson, Sundararajan, & Wojtach, 2014). The varying viewpoint likely gives the perceptual system better exposure to the three-dimensional identity of faces, allowing for a much more efficient representation of the true average. However all of these viewpoints seemed to have a frontal average viewpoint, and thus it is not known whether perceptual averaging of identity across varying viewpoints benefits from the global frontal orientation of the faces.

On the opposite account, identity invariant viewpoint extraction has also been reported (Florey, Clifford, Dakin, & Mareschal, 2016). Here, 16 faces, sampled randomly from four base identities, were presented to participants in varying orientations. Face viewpoints could have large standard deviations (viewpoints were more heterogeneous) or small standard deviations (viewpoints were more homogeneous). Using a 2AFC task, participants reported whether the average viewpoint direction was leftward or rightward (they did not continuously manipulate the direction to estimate the true average viewpoint). One task had participants reporting average
gaze direction and in another they reported average viewpoint direction. In both cases, participants could tell which direction average viewpoint or average gaze was, though Florey and colleagues found participant reports of viewpoint were more accurate than gaze direction.

On the surface, these results seem to demonstrate independence between viewpoint and identity extraction. Facial viewpoint is likely a lower level visual feature than identity, similar to gabor patch orientations described by Haberman, Brady, and Alvarez (2015). In extracting an average viewpoint, the visual system has to determine the angle of the stimuli in the visual field, and linearly combine them across that single dimension. Face identity, on the other hand requires much higher level abstraction across multiple dimensions such as the shape of the eyebrows, nose, mouth, and the geometric positioning of these facial landmarks (Nestor, Plaut, & Behrmann, 2016). With these higher order dimensions alone, image reconstruction from neural data is possible with greater than chance accuracy (Nestor et al., 2016).

When discussing the possible interaction between viewpoint and identity in face ensembles, the distinction between local and global viewpoints needs to be emphasized. The global viewpoint is the average direction of the faces. The local viewpoints are each individual face’s direction. In homogeneous viewpoints, the local viewpoint and global viewpoint are the same, where each individual face’s viewpoint is the mean viewpoint. For heterogeneous viewpoints, the local viewpoints are different from the mean viewpoint. Yamanashi Leib and colleagues (2014) describe accurate identity extraction when local viewpoint differs across a frontal global viewpoint. Despite the viewpoints of each face in a particular trial being different, the average direction converges frontally. Ensemble identity extraction may be preserved in this case because non-frontal viewpoints allow cognitive mechanisms to sample from different 3D angles. The most logical follow-up question, which this thesis aims to investigate, is what happens when the global viewpoint is also non-frontal? This question also arises in part because functional imaging has argued that true invariance does not occur, showing cortical activity changes following orientation changes that are independent from adaptation to repeated stimulus exposure (Andresen, Vinberg, & Grill-Spector, 2009).

In the interaction between global viewpoint and identity, there are two possible hypotheses. The first possibility is that identity extraction is affected by change in global viewpoint, meaning the two features interact. In this case, ensemble identity extraction could increase in efficiency or
decrease in efficiency as global viewpoints become non-frontal, with the second being the more likely outcome. The second possibility is that average identity extraction is not affected by change in global viewpoint, meaning they do not interact. This would mean that the cognitive system involved in face ensemble processing is similar to the independent higher- and lower-level processing model posed by Haberman, Brady, and Alvarez (2015); this is the independent model. If extracting an ensemble identity requires a linear combination of the contour shape features of the face, and not so much interpolating based on the holistic identity dimensions of each face (and this is a possibility based on prosopagnosics being able to process ensembles; Yamanashi Leib et al., 2012), non-frontal global viewpoints would not matter when extracting average identity.

In this thesis, I attempt to test these two hypotheses by manipulating identity and viewpoint of ensemble faces and compare how changes in one or both features affect the ability of participants to extract a summary identity. This thesis contains the results of two experiments and a brief proposal for a third. The first experiment is a proof-of-principle with computer generated stimuli. In it, I run a standard ensemble paradigm where participants report an average or single feature of either the identity or viewpoint. This is to ensure a valid setup of our novel paradigm. The second experiment is split into two parts. The first part (Experiment 2a) investigates how changes in global viewpoints, from 60° leftward to 60° rightward, affects the extraction of summary identity. Experiment 2b compliments 2a by investigating how varying the global viewpoint affects the extraction of a single-face identity from an ensemble.

The findings of the research contained in this thesis will provide insight into the underlying cognitive mechanisms of ensemble face processing. First, it will shed light on whether low and high level features (viewpoint and identity, respectively) interact with one another behaviourally, building on the previous model by Haberman, Brady, and Alvarez (2015). Second, comparisons of viewpoint and identity for both average and single item reports will demonstrate whether the dissociation between ensemble and single-item processing occurs for low-level (viewpoint) and high-level (identity) facial features. Third, I hope results assist in comparing mechanisms of face processing for single faces and face ensembles. The results will hopefully be built upon by future research in neuropsychology, cognitive neuroscience, and even computer vision as it pertains to the processing of face ensembles.
Chapter 2
Methods and Results of Experiments 1 and 2

2.1 General Methods

2.1.1 Participants

The experiments were approved by the University of Toronto Scarborough research ethics board. In total, there were 61 participants across Experiments 1, 2a, and 2b. Each participant volunteered in only one experiment and provided informed consent before data collection began. Participants were made up of first year undergraduate students participating in research for course credit, lab members and graduate students within the Department of Psychology, or students (undergraduate or graduate) from the general University of Toronto Scarborough campus recruited via poster and paid $10 CAD for participation.

All participants were screened with the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) to ensure adequate face processing ability. Data from participants who performed below 60% accuracy were discarded.

For the actual experiment, within condition outlier trials (±2.5 SD from their condition mean) were removed. If a single participant condition was more than 2.5 SD away from the mean performance across participants (above or below), data from that entire participant was dropped. If participants withdrew consent during an experiment, their data was also discarded. Finally, if participants showed strong fatigue, measured as a negative correlation between accuracy and the trial number, their data was removed.

Experiment 1 originally had 30 participants. Five were removed for having CFMT scores below 60%, one withdrew consent during the study, six were removed based on the outlier criteria described above, and one participant’s data was lost due to a computer malfunction. This left Experiment 1 with 17 total participants (mean age = 19, SD = 1.5; nine males, eight females).

Experiment 2a originally had 17 participants. Two people were removed for having a below threshold CFMT score, one person was removed due to showing strong fatigue (participant $r = -0.295$, average $r = -0.007$, $SD = 0.092$), and two were removed based on outlier criteria. This left Experiment 2a with 12 total participants (mean age 21, $SD = 2.2$; five males, seven females).

Experiment 2b originally had 14 participants. One participant withdrew their consent, one
participant had poor CFMT and outlier scores, and another participant was removed based on outlier criteria. This left Experiment 2b with 11 total participants (mean age = 22, SD = 4.3; seven males, four females).

2.1.2 Stimulus Generation

A portion of studies investigating face ensembles, including some mentioned above (e.g. de Fockert & Wolfenstein, 2009; Florey et al., 2015; Haberman & Whitney, 2007), used a forced-choice paradigm to investigate ensemble face perception. While this is a valid method to ascertain the cognitive perception of face ensembles, to directly compare identity and viewpoint features, a more precise measurement system was needed. Here, I generated a set of dynamic 3D faces which could differ in viewpoint or identity. Faces were generated using FaceGen Modeller Pro, version 3.5 (Singular Inversions, http://facegen.com/modeller.htm). FaceGen is capable of generating high fidelity 3D full-head images both at random and from imported face images. In this research, I used randomly generated faces and linearly morphed them to obtain a continuous selection of faces. FaceGen uses a parameterized method to linearly modify hundreds of facial features, making it efficient for use in generating morphs from different preset identities by manipulating feature sliders or weights between preset faces.

A large set of unique faces were initially randomly generated by FaceGen, and three that appeared most perceptually dissimilar were chosen as anchor faces. From there, I generated a set of linear identity morphs by weighing features between two anchor faces. The number of morphs generated between any two anchors was altered until there was a just noticeable difference between any two morphs, resulting in 24 morph identities between anchors 1 and 2, 2 and 3, and 3 and 1. The morphs were weighted linearly depending on how close they were to each anchor. With all morphs and anchors there were a total of 75 identity units (IUs) arranged circularly. Unlike viewpoint degrees, IUs are arbitrary labels, referring to the distance between any two identities. IU 1 refers to the first anchor face, IU 26 refers to the second anchor, and IU 51 refers to the third anchor. The IUs 2-25, 27-50, and 52-75 refer to each linear morph identity between the anchors (Figure 1a-b).

Since the 75 identities were rendered in 3D, a snapshot of each viewpoint could be taken for each identity. An image of each identity from -90° (left profile) to 90° (right profile), including 0° (frontal) was exported in .tif image format. Every other degree of viewpoint angle was skipped
as the just noticeable difference between viewpoints was every 2° change. This resulted in 91 viewpoints for each identity (Figure 1c), and crossing each identity with each viewpoint yielded a total of 6,825 face images. Unlike the identities, viewpoints were not circularly arranged.

All images were hue controlled, so that the only diagnostically discernable feature for identity was the shape of the faces. This was to ensure low-level features such as skin colour, eye colour, blemishes, and others could not be used to determine average identity. To maximize stimulus-background contrast, the faces were presented on a black background.

2.1.3 General Procedures

After informed consent, participants filled out a demographic questionnaire, and then sat in a darkened room, 60 cm away from a screen with a 1920x1080 resolution and 59.9 Hz refresh rate. Participants were given instruction on the experimental task. Before completing the main experiment, participants completed 5-10 acquisition trials followed by some practice trials. Acquisitions trials differed from practice trials in that participants matched a feature of one face displayed onscreen to a match-to face displayed on screen, allowing them to familiarize themselves with the keyboard controls. The number of each depended on which experiment they were in. After the main experiment, they filled out a subjective response questionnaire and were debriefed on the purpose of the experiment. Each participant took approximately 50 minutes to complete the whole experiment.

For all experiments, a single script with modifiable parameters based on the experimental demands was programmed in MATLAB (https://www.mathworks.com/), version 2015a, with Psychophysics version 3.0.13 (Brainard, 1997; http://psychtoolbox.org/). The number of trials differed for each experiment. Each trial contained four phases: (1) a fixation phase, (2) a display phase, (3) a reporting phase, and (4) an intertrial interval (ITI). Each trial began with the fixation phase which contained a white fixation cross at the center of the screen between 300 and 800 ms in steps of 100 ms to prevent anticipatory responses. After, during the display phase, an ensemble consisting of six faces which varied in identity, viewpoint, or both (depending on the experimental paradigm), appeared on screen for 400 ms. During the display, the fixation cross stayed onscreen. Participants were instructed to remain fixated at all times and not to divert their eyes to any of the faces. Faces were presented evenly spaced in a circular arrangement, three on the left and right of the vertical center of the screen. Each face was 160 pixels vertically,
subtending a 4.24° visual angle. The subtending width depended on the viewpoint of each face. The diameter of the distance to the middle of any two faces was 366 pixels, which subtended a 9.69° visual angle.

After 500 ms the cross disappeared and a single probe face appeared at the center of the screen. Using the arrow keys, participants pressed up-down to filter through identities and left-right to filter through viewpoints, allowing them to report the perceived average feature. This precision measurement, unlike using a forced choice paradigm, will provide an informative comparison on the potential interaction of viewpoint and identity processing. When the participant finished their selection, they pressed the space bar which initiated the 500 ms ITI where the white cross reappeared before the next trial began.

In each experiment, the mean viewpoint and identities were predetermined and the order of appearance was randomized. To ensure orthogonality of viewpoint and identity, every combination of mean viewpoint and identity was shown to participants. In each trial, the mean feature was not shown to participants, instead a strictly-controlled set of local features were selected and displayed to participants. For viewpoint feature, each local viewpoint was -20°, -12°, -4°, +4°, +12°, and +20° away from the mean viewpoint (negative refers to leftward and positive refers to rightward relative to participant’s view). For example, if the chosen mean viewpoint was 10°, the six local viewpoints would -10°, -2°, 6°, 14°, 22°, and 30°. For the identity feature, each local identity was -15 IUs, -9 IUs, -3 IUs, 3 IUs, 9 IUs, and 15 IUs away from the mean identity (negative refers to counterclockwise on the identity circle, positive refers to clockwise on the identity circle). In this case if the mean identity was IU 35 (a morph between Anchor 2 and 3), the local identities would consist of IUs 20, 26, 32, 38, 44, and 50. Because the identities are arranged circularly, they cycle around after IU 75, or before IU 1. For example if the mean IU was 70, the local IUs would be 55, 61, 67, 73, 4, and 10. The steps between each viewpoint and identity were chosen so that all the local features took up about 40% of the total feature space (Figure 1d-e).

During the reporting phase, and for the feature that participants had to report, the starting feature of the probe was a random feature within the peripheral feature space (Figure 1d-e). This was to prevent the probe having any accidental familiarity to the ensemble set. When features were not reported, they were locked to the average feature in the reporting probe to maintain perceptual
consistency. For a non-reported feature, that feature in the reporting probe was locked to the mean of that feature to keep perceptual consistently with the ensemble set. In the event a participant was reporting a single feature (Experiment 1 or 2b) the unattended feature was not set as the mean of the unattended feature, but to the chosen single-face feature. Again this was to ensure maximal perceptual consistency.

During the main experiment, participants were given five 20-second breaks, each after completing 16%, 33%, 50%, 67%, and 84% of trials. The number of trials in each experiment varies, which alters how many trials take place between each break.

After each reporting phase during practice trials, participants were shown two faces simultaneously, one showing their reported feature and the other showing the true reported feature. This feedback allowed participants to better understand the experimental task and the error of their response. During the actual experiment, this feedback was not given.

Parameters of the ensemble display script were modified to fit the research question for each experiment. The later sections will describe the specific methods and procedures that have been modified from the heretofore described general methods.

2.1.4 General Statistical Analyses

Statistical analyses were conducted using either MATLAB, or IBM SPSS (version 24.0.0.0). All analyses were compared at an alpha of .05 using two-tails when necessary. For multiple comparisons, Bonferroni corrections were conducted to ensure < .05 family wise error rate. For within-subjects comparisons, Mauchly’s test of sphericity was applied to the data. When violated, the Greenhouse-Geisser corrected p value was reported. For between-subjects comparisons, homogeneity of variance was also tested. For t tests, measures of effect size were reported using Cohen’s d, where the pooled standard deviation was calculated as the square root of the summed group variance divided by 2.

2.2 Experiment 1: Comparison of Individual and Summary Extraction of Viewpoint and Identity from Face Ensembles

One purpose of Experiment 1 was to establish the validity of the novel paradigm created for this research by ensuring it replicates the general ensemble model. Specifically, they would be more
accurate when reporting an average feature than a single feature. Here, participants viewed face ensembles with differing viewpoints and identities. In each trial, the feature that participants did not report was held constant (i.e. local features were homogeneous, not varying from the mean). A second purpose for Experiment 1 is to directly compare accuracy across high-level facial features (identity) and low-level facial features (identity).

2.2.1 Methods and Procedure

There were ten acquisition trials. In these trials, participants saw two faces on screen, a probe face and a face whose feature (i.e. identity or viewpoint) could be varied to match the probe face. The identity and viewpoints of the faces to be matched were random. In five acquisition trials participants matched the probe viewpoint, and in the other five they matched the probe identity. These quick trials were to familiarize participants with the face stimuli, the method of changing viewpoint and identity continuously, and accustom them to the keyboard controls.

After the acquisition trials were the practice trials. For Experiment 1, there were 24 practice trials, each giving feedback after participant response by showing their selected face along with the actual feature for comparison. Here, a combination of three mean identities and three viewpoints were used. For six trials each participants reported the average identity, single identity, average viewpoint, or single viewpoint. The local feature that they were not reporting was homogenous. For example, if participants were reporting identity (either average or a single face), all local viewpoints and the mean viewpoint would have the same value and vice-versa for identity when participants reported viewpoint. Again, this was to prevent attended and unattended local features interacting with one another in Experiment 1.

After the practice trials, there were 216 experimental trials. These were split into 54 trials each where participants reported average identity, single identity, average viewpoint, or single viewpoint. When reporting identity, there were nine preset mean identities across three preset viewpoints, resulting in 27 total unique combinations, and each combination was presented twice. The preset mean identities were divided so that three were between each anchor set. When reporting viewpoint, the number of preset means for each feature were switched. Specifically, there were three preset mean identities across nine preset mean viewpoints, with 27 unique combinations that were repeated twice. For viewpoints, one of the preset means was 0° (frontal), with four preset mean viewpoints to the left and to the right in steps of 7°. The order of all trials
in the experiment was random. Breaks took place every 36 trials. After the display, during the reporting phase, the ensemble disappeared and the fixation cross changed colour to indicate which feature would be reported: green for identity, red for viewpoint (Figure 2).

Because each face ensemble had one feature with heterogeneous local values, and another feature with homogenous local values, participants could anticipate whether they would report identity or viewpoint. This was intentional to prime the visual system to focus on one feature instead of the other. However, participants were not pre-cued to whether they would report that feature as the average of the set, or from a single randomly-selected face. They were still cued with either the green (identity) or the red (viewpoint) fixation cross after the display phase and before the report phase. When reporting single features, the program chose a random face from the set as the true feature (either identity or viewpoint), and during the reporting phase a white oval appeared onscreen where that chosen face was previously displayed (see Figure 2). The presence of the oval indicated to participants they were reporting the single feature of that face and not an average of the set.

2.2.2 Results and Discussion

Comparison of Preset Identities and Viewpoints. The first thing I wanted to measure is if some morph identities that made up the preset means were harder to report than others. Recall that between each combination of Anchors 1 and 2, 2 and 3, and 3 and 1, there were three morph identities that were selected as the preset means in Experiment 1. In the event that reporting some morph identities were associated with larger error than others, it may indicate a floor-effect where dissociating between certain morphs was too difficult, or that two anchors may be too perceptually similar to one another to dissociate different morphs between them. To analyze this I conducted a 3 (Anchor Set: 1 to 2, 2 to 3, or 3 to 1 morph anchors) by 3 (Morph Location: closer to the first anchor, middle of both anchors, or close to second anchor) within-subjects ANOVA comparing the error in absolute IUs that the participant’s identity selection was from the true identity (Figure 3a). There was no difference in error between the closeness of any morph identity to either anchor, reflected in a non-significant main effect of Morph Location ($F_{2,32} = 0.67, p = .518$). The main effect of Anchor Set was close to, but did not reach, significance ($F_{2,32} = 2.82, p = .075$). Participant error for morphs between Anchors 1 and 2 ($M = 17.25, SEM = 0.718$) was a little higher than the error for morphs between Anchors 2 and 3 ($M =
14.42, $SEM = 1.11$), and Anchors 3 and 1 ($M = 15.36, SEM = 0.69$). It is therefore likely that morphs between some anchor sets may be perceptually similar than other morphs. However this does not confound the results because of the orthogonal arrangement of all identities and viewpoints.

I also investigated whether certain viewpoints were more difficult to average than others. In this case I measured both single and average viewpoint in a 2 (Task: report average or single) by 9 (Global Viewpoint at -56°, -42°, -28°, -14°, 0°, 14°, 28°, 42°, or 56°) within-subjects ANOVA (Figure 3b). There was a significant main effect of Task ($F_{1,16} = 72.87, p < .001$, partial $\eta^2 = .820$) and Global Viewpoint ($F_{8,128} = 16.36, p = .038$, partial $\eta^2 = .117$), but no significant interaction ($F_{8,128} = 0.76, p = .637$). Post hoc tests were calculated with a $p = .025$ Bonferroni corrected alpha. There was a significant effect of viewpoint at average task ($F_{8,128} = 2.52, p = .014$, partial $\eta^2 = .136$), but not for viewpoint at single task ($F_{8,128} = 0.99, p = .447$). Since there was an apparent curvilinear appearance of viewpoint extraction, I fitted a quadratic curve to average and single viewpoint extraction and tested the significance with a polynomial contrast. There was a significant quadratic fit of average viewpoint extraction ($F_{1,16} = 4.78, p = .044$, partial $\eta^2 = .230$) and single viewpoint extraction ($F_{1,16} = 4.53, p = .049$, partial $\eta^2 = .221$). This suggests that the ability to extract either an average or single viewpoint diminishes as global viewpoints become progressively less frontal. Note that both these findings, while below the .05 cutoff value, are no longer considered significant after a $p < .025$ Bonferroni correction. Nevertheless they both still exhibit a strong trend with a decent portion of the variance explained by the curve of each polynomial.

**Accuracy of Average and Single Feature Extraction.** Because there are a different number of total identities and viewpoint degrees, I cannot compare these two features by simply stating how many steps away a participant’s report was from the true feature, in either IUs or degrees. Instead I will be using the mean percent deviation (MPD) by taking the absolute number of steps between the participant’s report and the true feature, and dividing over the total range of that feature. This will allow me to compare viewpoint and identity performance by contrasting the percent error of participants’ reports.

Identity and viewpoint reports from participants were transformed into MPD and analyzed with a 2 (Feature: report identity or viewpoint) by 2 (Task: report average or single) within-subjects
ANOVA (Figure 4a). Omnibus results revealed a significant main effect of Feature ($F_{1,16} = 1258.71, p < .001, \text{partial } \eta^2 = .987$), Task ($F_{1,16} = 17.78, p = .001, \text{partial } \eta^2 = .526$), and a significant interaction ($F_{1,16} = 14.91, p = .001, \text{partial } \eta^2 = .482$). Post hoc analyses were conducted to investigate error differences when extracting an average or single feature. There was no significant difference in error between extracting average and single identity (average identity = 20.90 MPD, single identity = 21.56 MPD; $t_{16} = 0.86, p = .400$), though there was a significant difference in error difference between average and single viewpoint (average viewpoint = 6.32 MPD, single viewpoint = 9.79 MPD; $t_{16} = 8.69, p < .001, d = 2.5$).

The efficiency at extracting average or single identity or viewpoint was correlated among participants using a Pearson product correlation; the results are presented in Table 1. The resulting correlation coefficient was transformed into a $t$ statistic ($df = 15$) for significance testing. All correlations were non-significant (all $p$s $> .230$), except one which was marginally significant when average identity was compared with average viewpoint ($r = .463, p = .061$). This suggests that extracting an average identity is dissociable from single identity extraction, but there may be some cognitive overlap with extracting average identity and average viewpoint. Of course, more participants are needed to validate this finding as the correlation was only marginally significant.

**Speed of Processing For Average and Single Feature Extraction.** I calculated both response onset (RO) as the time it takes participants to initiate a keyboard input once the reporting phase starts, and selection time (ST) as the time between RO and the completion of probe feature. The total time (TT) is the sum of RO and ST. The results are plotted in Figure 4b. All three of these metrics provide an indication of speed of processing, and I evaluated them using a 2 (Feature: identity or viewpoint) by 2 (Task: report average or single) within-subjects ANOVA. For total TT, the significant main effect of Feature revealed that participants were faster at reporting viewpoint than identity (mean identity TT = 3,701 ms, mean viewpoint TT = 2,541 ms; $F_{1,16} = 38.26, p < .001, \text{partial } \eta^2 = .705$). The main effect of Task was not significant ($F_{1,16} = 1.11, p = .308$), nor was there an interaction between Feature and Task ($F_{1,16} = 0.02, p = .897$). The finding that participants reported viewpoint faster than identity even though there are more possible viewpoints to select from (91 compared to 75 identities) indicates faster overall processing of viewpoint than identity, likely because it is a lower level visual feature.
This interpretation is also supported by the finding that participants were faster to begin making their selection (RO) for viewpoint than identity (mean identity RO = 845 ms, mean viewpoint RO = 724 ms; significant main effect of Feature: $F_{1,16} = 6.50, p = .021$, partial $\eta^2 = .289$). The RO likely reflects some cognitive preprocessing taking place, enough for participants to have an initial estimate of the feature they are reporting. Similar to the results with TT, here there was no difference in RO for reporting average or single features (main effect of Task: $F_{1,16} = 0.65, p = .432$) nor was there an interaction ($F_{1,16} = 0.48, p = .498$). For ST, participants were also significantly faster at completing the viewpoint feature selection than the identity feature selection (mean viewpoint ST = 1,817 ms, mean identity ST = 2,856 ms; significant main effect of Feature: $F_{1,16} = 29.68, p < .001$, partial $\eta^2 = .650$). Again, the main effect of Task for ST ($F_{1,16} = 0.34, p = .571$) and the interaction ($F_{1,16} = 0.01, p = .952$) were both non-significant.

**Average Identity Influencing Single Identity Perception.** While there was no difference in MPD between single and average identity extraction (Figure 4a), it may be possible that the global identity was biasing single-identity extraction. This global-bias is consistent with previous research (Hochstein, Pavlovskaya, Bonneh, & Soroker, 2015). One empirical study using circles found that the remembered size of individual circles was biased towards the mean size of the set (Brady & Alvarez, 2011). Here I report MPD for single identity or single viewpoint extraction when referenced to the true indicated single face or when referenced to the mean of that feature. Results were analyzed by a 2 (Feature: identity or viewpoint) by 2 (Reference: referenced to the true single feature, or referenced to the mean of that feature) within-subjects ANOVA (Figure 5). There was a significant main effect of Feature, with viewpoint having less overall MPD than identity ($F_{1,16} = 479.23, p < .001$, partial $\eta^2 = .968$). There was also a significant main effect of Reference, with less MPD when the reported single face feature was referenced to the average feature than the true feature from the single face ($F_{1,16} = 28.88, p < .001$, partial $\eta^2 = .643$). What this latter finding suggests is that the reported single feature for both identity and viewpoint is more biased by the average of the set than by the actual single face feature that they were meant to report. Next, there was a near-significant Feature-by-Reference interaction ($F_{1,16} = 4.13, p = .059$), suggesting that the identity feature may be more biased by the mean than the viewpoint feature.

**Experiment 1 Discussion.** Experiment 1 investigated the basic ensemble processing paradigm as it applies to the extraction of identity and viewpoint features. Here I report that participants were
more efficient (there was less MPD) at extracting viewpoint than identity. Within identity, there was no difference between average or single identity extraction, and within viewpoint there was a difference between average or single viewpoint extraction. This points to a possible dissociative mechanism for single and average viewpoint processing, but not for single and average identity processing. The standard ensemble model assumes a greater efficiency for extracting a global summary than any individual item in a set. The absence of this finding for facial identity has been noted previously (Haberman, Brady, & Alvarez, 2015) and speaks to the idea that facial identity may be a special exception to the rule. The equal efficiency in extracting an average and single identity is consistent with the proposal that processing a summary identity may not preclude or negate the extraction of individual identity (Neumann, Ng, Rhodes, & Palermo, 2017).

For viewpoint extraction, there was an increase in error as global viewpoints became progressively less frontal (Figure 3b). Because the six local viewpoints were the same distance relative to the global viewpoint, the finding that non-frontal global viewpoints result in decreased efficiency (larger MPD) suggests that viewpoint processing is benefited by a frontal angle. It is possible that this result arises because there are fewer neurons dedicated to processing non-frontal faces (Ramírez, Chichy, Allefeld, & Haynes, 2014). Another explanation is that there are fewer areas of visual space that are diagnostic to viewpoint when faces are non-frontal. Future research will hopefully investigate these two possibilities, among others. One method is to investigate ensemble viewpoint extraction for non-face stimuli, such as cars. The other is to replicate these results with a prosopagnosic patient group. Those investigations should help shed light as to whether viewpoint extraction requires some identity processing. In general, I replicated previous findings by Florey and colleagues (2016), that observers can extract a summary representation of viewpoint, and added that this efficiency decreases as viewpoints progress non-frontally.

Participants took almost 100 ms longer to begin reporting identity than viewpoint (Figure 4b) regardless of reporting the average or single feature. This suggests two things. The first is that viewpoint is extracted faster than identity. Before the reporting phase, participants are primed with a green or red fixation cross to tell them which feature they will report, and so these results cannot be explained by uncertainty or personal preference to which feature will be reported.
Second, these results suggest information about features for single faces are extracted either in parallel with summary feature extraction, or are extracted from the summary feature. The equal RO for single or average extraction is consistent with both of these interpretations. However, I believe the latter to be more likely. First, previous research has found that reporting single-face identity is biased by the mean identity (de Fockert & Wolfenstein, 2009). Second, the evidence presented in a review of ensemble processing suggest global statistical summaries act as a baseline for perceptual processing of individual items (Hochstein et al., 2015). Third, the findings here where participant’s reports of any single feature is biased by the mean feature (Figure 5) is consistent with the idea that summary extraction informs single extraction direction.

These interpretations are not conclusive however, and further research is warranted to investigate the influence of single-to-summary or summary-to-single biases in ensemble processing in both face and non-face ensembles. I also suggested here that face viewpoint extraction has some reliance on a face-processing network and is not merely a geometric averaging of orientation. This can be expanded upon by future research, adding to the understanding of the cognitive mechanisms involved in summary viewpoint extraction. This interpretation will be elaborated on in the general discussion.

2.3 Experiment 2a: Viewpoint Sensitivity in the Extraction of Summary Identity from Face Ensembles

In this experiment, I explore a possible interaction between viewpoint and identity features by investigation how the extraction of a summary identity is affected by changes in global viewpoint. In Experiment 1, when participants were extracting identity the local viewpoints were homogenous with the global viewpoint. The difference in the next experiment is that both global and local viewpoints will vary, meaning participants will have to sample identity from faces that each have different local viewpoints and identities instead of just different local identities. This is a direct extension of Yamanashi Leib and colleague’s (2014) research on summary extraction across varying viewpoints, but with global viewpoints changing as well as local viewpoints. It is also more consistent with real-world exposure to faces as we do not always view crowds where the average viewpoint is towards us. This also affords the ability to investigate precisely how varying viewpoints interact with identity extraction by comparing these results with Experiment 1 where local viewpoint values do not change. A total of 12 new participants volunteered for
2.3.1 Experiment 2 Methods and Procedure

All parameters described in the general methods remained the same. In this experiment, participants reported only average identity. In all trials, both local identity and local viewpoint differed from their global feature. There were 13 preset mean viewpoints covering a wider-range of viewpoints than Experiment 1: -60°, -50°, -40°, -30°, -20°, -10°, 0° (frontal), 10°, 20°, 30°, 40°, 50°, and 60°. There were also 15 preset mean identities (five between each anchor), all of which were five IUs apart. There were a total of 195 trials, which is the orthogonal combination of all preset identities and viewpoints. Participants were given five 20 second breaks every 32-33 trials. Like the previous experiment, there were acquisition trials consisting of five identity-matches, and 24 practice trials made of eight preset identities and five preset viewpoints (three viewpoints were shown six times, two were shown three times).

2.3.2 Results and Discussion

For Experiment 2a (and later 2b), data was not transformed into MPD because identity was the only feature being analyzed. All data was analyzed in SPSS in a one-way within-subjects ANOVA comparing average identity extraction across each of the 13 global viewpoints.

**Comparison of Preset Mean Identities.** Like Experiment 1, I analyzed whether any of the 15 preset mean identities were harder to extract than others, regardless of viewpoint. All preset mean identities were arranged in a 3 (Anchor Set: between Anchor 1 and 2, 2 and 3, or 3 and 1) by 5 (Morph Location: closest to the first anchor, close to the first anchor, middle of both anchors, close to second anchor, or closest to second anchor) within-subjects factorial ANOVA and the results of this analysis are plotted in Figure 6. Data shows the distance in IUs that participant’s report of average identity was from the true average identity. Here, there was no effect of the relative closeness to either the first or second anchor for any morph identity (non-significant main effect of Morph Location: $F_{4,44} = 1.74, p = .159$), but there was an effect of Anchor Set ($F_{2,22} = 5.89, p = .009$, partial $\eta^2 = .349$). There was also no significant interaction between these factors ($F_{8,88} = 0.93, p = .497$). Post hoc analyses revealed that the second anchor set (morphs between Anchors 2 and 3) had significantly less error (participant reports of average identity were fewer IUs from the true mean identity) than the first Anchor Set (morphs between Anchors 1 and 2: mean difference = 1.9 IUs, $p = .006$) and the third Anchor Set (morphs between Anchors 3 and 1: mean difference = 2.4 IUs, $p = .045$). It is entirely possible that discriminability
between any two anchor faces is not perceptually equal. Since Experiment 2 contains more morph identities between each anchor set, any perceptual ambiguity is more likely to emerge. Nevertheless, I do not believe these differences call into question the validity of subsequent analyses for two reasons. First, even though statistically significant, the discrepancy in error for reporting morphs between any anchor set is quite small (Figure 6), ranging from 11 IUs for the first identity in anchor set 2-3, to 15.9 IUs from the middle identity in anchor set 3-1. That gives a range of 4.9 IUs, which is smaller than the distance between any two local identities. There are no identities or groups of identities that show striking differences in patterns of error than others. Second, any ease or difficulty in differentiating between certain identities is evenly spread across all viewpoint ranges. That is, every preset mean identity is shown for every preset mean viewpoint.

**Identity Extraction Across Varying Viewpoints.** The main purpose of Experiment 2a was to investigate summary identity extraction across a range of non-frontal global viewpoints. Here, error in summary identity (average IUs away from the mean identity) at every global viewpoint was analyzed as a one-way within-subjects ANOVA (Figure 7, blue line). While the one-way ANOVA was non-significant ($F_{12,132} = 1.213, p = .281$), polynomial curve fitting revealed a significant fit for a quadratic function ($F_{1,11} = 5.12, p = .045$, partial $\eta^2 = .318$), indicating a parabolic increase in IUs as global viewpoints deviated non-frontally.

**Speed of Processing for Average Identity Extraction.** Like Experiment 1, I analyzed the timing of identity extraction by examining RO, ST, and TT (Figure 8a). Here, the effect of global viewpoint on TT when extracting average identity was marginally significant in a one-way within-subjects ANOVA ($F_{12,132} = 1.71, p = .072$). Polynomial curve fitting revealed a significant quadratic function ($F_{1,11} = 16.94, p = .002$, partial $\eta^2 = .606$), with high TT for frontal faces and a steep drop off as faces became progressively non-frontal. This parabolic curve in TT was driven by both ST and RO, as two individual quadratic fits were significant for both ST (one-way within subjects ANOVA $F_{12,132} = 1.96, p = .033$, partial $\eta^2 = .151$; quadratic curve fit $F_{1,11} = 21.51, p < .001$, partial $\eta^2 = .662$) and RO (one-way within subjects ANOVA $F_{12,132} = 2.17, p = .016$, partial $\eta^2 = .165$; quadratic curve fit $F_{1,11} = 7.519, p = .019$, partial $\eta^2 = .406$). The curves for RO and ST were in opposite directions, with an increase in time for RO and a decrease in time for ST as face became non-frontal.
**Experiment 2a Discussion.** The main finding of this experiment is that the efficiency of average identity extraction suffers as global viewpoints become progressively non-frontal (Figure 7). This is consistent with previous research on viewpoint perception in objects (Andresen et al., 2009) and single faces (Caharel, Collet, & Rossion, 2015). It is likely that in ensemble identity extraction at least some part of the face-processing network is implicated and it is not merely a linear interpolation of surface features and shape. If it were, I would not have expected a significant parabolic distribution in error (increase IU for non-frontal faces). Whether this face-processing network is the same as for single-faces, or an ensemble face network distinct from single-faces, paralleling mechanisms for the processing of single versus object ensembles (Cant et al., 2015), is an important question for future research.

RO was quicker for face ensembles oriented in a frontal global viewpoint, but ST took longer (Figure 8a). It is reasonable to assume, as mentioned in Experiment 1, that before participants begin their input at least some initial processing has taken place. Identities oriented frontally may be more efficiently processed due to the abundance of frontally-tuned face processing neurons in visual cortex (Ramírez et al., 2014). The finding that it takes longer for participants to complete their selection (longer ST) may speak to the visual system having more identity-related information due to more efficient processing, allowing participants to be more thorough in their report of average identity by fine-tuning their selection. This is supported from the findings in both Experiment 1, and from the finding that summary identity extraction in Experiment 2a is more accurate for frontal faces than non-frontal faces.

These findings, while behavioural, may indeed speak to potential neural mechanisms of ensemble face processing. However, future research, using a combination of temporally-sensitive (EEG) and anatomically-sensitive (fMRI) techniques is still needed to establish the neural mechanisms taking place. Specifically, if lower RO for frontal faces really does reflect faster access to identity information, this should be evident from face-related ERP signatures. Second, if processing of ensemble identity utilizes the same face networks for single-face identity processing, this can be confirmed through comparative activation of cortical regions through fMRI. Perhaps in addition to domain-specific cortical activation there is also activation in scene selective cortex, paralleling differences in activation for single objects (Cant & Goodale, 2007; Cant et al., 2009) and object ensembles (Cant & Xu, 2017).
2.4 Experiment 2b: Viewpoint Invariance in the Extraction of Single Face Identity from Face Ensembles

After the results of Experiment 2a, I was curious if the resulting effect would be replicated when participants were instructed to extract a single identity from face ensembles, so I ran a second part of Experiment 2.

2.4.1 Methods and Procedure

All experimental procedures were the same as Experiment 2a, including parameters for both practice trials (24 trials) and experimental trials (195 trials). The only difference here was that instead of participants always reporting the summary identity of the faces, they were now instructed to always report the identity of a single, randomly selected face.

2.4.2 Results and Discussion

As mentioned previously in Experiment 2a, data was not transformed into MPD because identity was the only feature being analyzed. All data was analyzed in SPSS in a one-way within-subject ANOVA comparing single identity extraction across each of the 13 global viewpoints.

**Single Identity Across non-Frontal Viewpoints.** Figure 7 shows the results of single-identity extraction across multiple global viewpoints. Here, there was no significant effect of global viewpoint on single identity extraction \((F_{12,120} = 0.94, p = .512)\), nor was the fitted quadratic curve significant \((F_{1,10} = 0.01, p = .923)\). This indicates that, unlike reporting average identity, reporting a single identity from face ensembles is not affected by changes in global viewpoint.

**Average Identity Influencing Single Identity Perception.** In Experiment 1 I assessed whether single feature extraction was influenced by the global feature. Similarly, here I analyzed whether extraction of a single face identity was influenced by the mean identity across varying viewpoints (Figure 9). This was analyzed as a 2 (Reference: to true single identity, or to the average of the set) by 13 (all viewpoints) within-subjects ANOVA. The main effects of Reference \((F_{1,10} = 0.63, p = .444)\) and Viewpoint \((F_{12,120} = 0.75, p = .699)\) was not significant, nor was there an interaction \((F_{12,120} = 1.45, p = .153)\). This suggests that when reporting single identity in Experiment 2b, there is an equal contribution of the actual single identity from the set, and the average of the set.
**Speed of Processing for Single Identity Extraction.** Like Experiments 1 and 2a, I analyzed the timing of identity extraction by examining RO, ST, and TT. The results are plotted in Figure 8b. For overall TT there was no significant difference across the 13 preset mean viewpoints \((F_{12,120} = 0.53, p = .889)\), nor did polynomial curve fitting reveal a significant fit of the quadratic function \((F_{1,10} < 0.01, p = .994)\). For both RO and ST, there was no significant difference across the 13 preset mean viewpoints \((\text{RO } F_{12,120} = 1.01, p = .443; \text{ ST } F_{12,120} = 0.84, p = .610)\), nor did polynomial curve fitting reveal a significant fit of the quadratic function \((\text{RO contrast } F_{1,10} = 0.64, p = .441; \text{ ST contrast } F_{1,10} = 0.06, p = .820)\). This seems to suggest that, unlike Experiment 2a, the time it took to initiate or complete the response for single-face identity was not influenced by global changes in viewpoints, demonstrating invariance.

**Comparison to Experiment 2a.** There was no difference in the error for reporting single versus average identity in Experiment 1 (Figure 4). However, in Experiment 2, I report a between-subject difference in IUs for the extraction of average identity (2a) and single (2b), which can be visualized in Figure 7. To statistically compare this difference, the data for Experiments 2a and 2b were combined and analyzed in a 2 (Between Subjects Task: report average or single identity) by 13 (Within Subjects Viewpoint: all 13 viewpoints) mixed-factorial ANOVA. The main effect of Viewpoint was no longer significant with the combined data \((F_{13,252} = 0.66, p = .789)\). This was likely due to viewpoint-dependent differences in average identity extraction being diluted by the non-difference among single identity extraction. There was also no significant interaction effect \((F_{12,252} = 1.47, p = .137)\). The main effect of Task was significant \((F_{1,21} = 73.08, p < .001, \text{ partial } \eta^2 = .777)\), providing empirical support for the observation that participants made more errors when reporting single identity compared with mean identity. These findings are in contrast to what I observed in Experiment 1 (no dissociation between single and mean facial-identity processing as indexed by error).

Comparisons of TT in Experiments 2a and 2b was also done (Figure 8), but this analysis was collapsed across viewpoints. For RO, participants had a faster onset when reporting average identity compared with single identity (Experiment 2a mean = 485 ms, \(SD = 118\) ms; Experiment 2b mean = 683 ms, \(SD = 184\) ms; \(t_{21} = 3.10, p = .006)\), but for ST participants took longer to report an average identity than a single identity (Experiment 2a mean = 3,957 ms, \(SD = 1,325\) ms; Experiment 2b mean = 2,602 ms, \(SD = 1,214\) ms; \(t_{21} = 2.55, p = .019)\). In summary, when
averaging, compared to extracting a single identity, participants were faster to initiate an onset but took longer to complete the selection.

**Experiment 2b Discussion.** The most obvious difference between Experiment 2 and Experiment 1 is that here I report a dissociation between the processing of single and average facial identity that was not apparent in the first experiment. This is not necessarily a contradictory finding. Recall in Experiment 1, local viewpoints were homogenous with the global viewpoint when participants extracted identity. Participants can efficiently extract a summary identity across varying viewpoints (Yamanashi Leib et al., 2014), and results from Experiment 1 suggest that average identity biases the extraction of single identity – and this is noted in other research (Brady & Alvarez, 2011; de Fockert & Wolfenstein, 2009). When comparing average identity extraction for when local viewpoints were homogenous (Experiment 1) compared to when local viewpoints were heterogeneous (Experiment 2a), I actually found that participants were better at extracting summary identity when local viewpoints differed (Experiment 1 mean = 15.68 IUs, $SD = 2.06$ IUs; Experiment 2a mean = 13.27 IUs, $SD = 1.49$ IUs; $t_{27} = 3.45, p = .002$). When comparing single identity extraction when faces were homogenous (Experiment 1) versus heterogeneous (Experiment 2b) I found the opposite results, in that participants were better at extracting a single identity when local viewpoints did not differ (Experiment 1 mean = 16.17 IUs, $SD = 0.97$ IUs; Experiment 2b mean = 17.60 IUs, $SD = 0.81$ IUs; $t_{26} = 4.03, p < .001$).

From these findings, it can be interpreted that when local viewpoints differ, the visual system can construct a better three-dimensional representation of the average identity. This best explains the more efficient summary extraction found for Experiment 2a than Experiment 1. For single identity extraction, because the local viewpoint is different from the global (mean) viewpoint, it may be more difficult to discount interactions of viewpoint, resulting in the discrepancy seen here where single identity extraction was worse for Experiment 2b than Experiment 1. It would also explain why in Experiment 2b single identity reports were evenly weighed between the true single face and the set average, but in Experiment 1 reports of single face reports were much closer to the mean face. In Experiment 2b, the visual system cannot be as reliant on the average face identity when estimating the single face identity because of the perceptual ambiguity that comes with non-converging viewpoints. Despite consistency with the present results, this latter interpretation is speculative and warrants future research to better elucidate.
Of course there are still alternative interpretations. For one, Experiment 1 was a mixed-trial design whereas Experiment 2 was a block design. One possible alternative explanation is a serial dependency or task switching phenomenon affecting participant’s performance in Experiment 1, but not Experiment 2. Given previous research which reported consistent findings using a block design (Haberman, Brady, & Alvarez, 2015), I do not think this is likely but additional research is still warranted to test this interpretation. For example, a third experiment with a fully within-subjects design could test average identity extraction in a mixed-block design where in some trials local viewpoints are homogenous and in others they are not.

In comparing Experiments 2a and 2b, it appears that there are some dissociable processes between the extraction of average and single identity, seeing how changes in both local and global viewpoint affect participant’s reports. It also appears that changes in viewpoints interfere with the ability to extract a summary identity which would not be expected if there were two independent systems for processing high and low visual feature ensembles (Haberman, Brady, & Alvarez, 2015). In the next chapter I will consolidate all three experiments, discussing possible interpretations of the data, how the results fit with what is known from current research, and its implications for future studies.
Chapter 3
General Discussion and Future Directions

The purpose of my thesis was to investigate the cognitive mechanisms mediating the extraction of features (i.e. identity and viewpoint) from face ensembles. The importance of this research is predicated by the social role faces play in our daily lives, both single faces encountered in situations such as one-on-one conversations, and face ensembles encountered in interactions with large crowds.

2.5 Highlights of General Findings

Experiment 1 investigated both summary and single-feature extraction for face identity and viewpoint, always holding one feature constant when the other was extracted (i.e. there were no local deviations for the unattended feature). Experiment 2a investigated summary identity extraction over varying viewpoints, and Experiment 2b was the same except with single-face extraction.

In Experiment 1, I found that extraction of a summary or single viewpoint is affected based on the global viewpoint (Figure 3b). As global viewpoint was progressively less-frontal, there was a curvilinear increase in error for both average and single viewpoint reports. Participants also had more error for extracting a single viewpoint than an average viewpoint (this dissociation was not seen in identity), and in general they were much faster at extracting viewpoint than identity (Figure 4a). There was also faster onset for reporting viewpoint than identity (Figure 4b). I also found that when reporting both single-features, the report was closer to the mean of the set than the actual single face feature (Figure 5).

In Experiment 2, participants showed a similar curvilinear increase in error as average face identity extraction occurred across progressively non-frontal viewpoints, but in Experiment 2b this was not seen in single faces, which contrast with the results in Experiment 1 (Figure 7). This is likely because local viewpoints were also varying in Experiment 2b, making it difficult to use the mean face as a reference face. Of course this assumes that the processing of features from single items is driven from the summary representation, a view consistent with the general understanding of global-local processing in the visual system (Hochstein et al., 2015). Report onset delay in Experiment 2 also supports this interpretation as participants took longer to initiate
a response when reporting single identity than reporting average identity (Figure 8). This could be because an individual identity percept needs to be drawn from the summary percept, increasing the duration to initiate a response.

Next, while it’s been previously reported that participants can extract a summary identity across varying viewpoints (Yamanashi Leib et al., 2014), the results comparing Experiments 1 and 2 together are the first to suggest that participants are more efficient when local viewpoints vary, likely because of enhanced three-dimensional sampling space. Furthermore, these are the first results to find that viewpoint invariance is based on local variations in viewpoint, not global, as summary identity extraction increased in error when mean viewpoints were non-frontal.

### 2.6 Implications for Future Research

The general ensemble processing model claims that extraction of a summary feature is more efficient or dominates the processing of features for individual items. This is often seen as less error when reporting the average feature than any one feature (e.g. Ariely, 2001), or claiming the mean feature was also a set member (de Fockert & Wolfenstein, 2009). In Experiment 1, where local changes in unattended features were controlled for, I found that this phenomenon holds true for viewpoint but not identity. Other research has found single-average dissociations in non-identity face-related features such as emotion (Haberman & Whitney, 2009), while the lack of a single-average dissociation for identity has been previously noted alongside the findings I present in Experiment 1 (Haberman, Brady, & Alvarez, 2015; Neumann, Schweinberger, & Berton, 2013; Neumann et al., 2017).

Recently, Neumann and colleagues (2017) found that when the ensemble display duration was increased beyond 1.5 seconds, the ability to extract an individual identity began to surpass that of average identity, even in a set size of eight faces. These results, especially in the context of the present research, appear puzzling. However, the social importance of face identity cannot be underestimated and older research has found that the capacity limitation of VWM may not actually apply to faces (Curby & Gauthier, 2007; Wong, Peterson, & Thompson, 2008), unlike lower level visual features such as orientation (Jiang, Shim, & Makovski, 2008). This capacity advantage in short-term memory is also true for other categories of acquired expertise, such as with car experts (Curby, Glazek, & Gauthier, 2009).
If single-face processing is not as affected by capacity limitations like other visual features, this may explain why single-face extraction was as efficient as summary face extraction with six faces present. How are faces (and expertise) not as limited as other objects? This phenomenon may be explained by ensemble processing. In Experiment 1, single face reports were closer to the mean identity than they were the true identity. Perhaps the increased capacity to hold individual face identities is driven by the global storage of summary features. If the visual system compiles a perceptual average of faces, then less individual features of each face need to be stored or individual features can be stored as a deviation from the average features. When perceptually reconstructing a single face, the visual system may use the average face as a preset and modify it with stored features about individual faces. This can take place simultaneously, or serially where exposure to new faces are averaged over time (Liberman, Fischer, & Whitney, 2014). If this were the case, exposure to certain faces may dictate how well we can store information from new faces depending on how similar or distant they are from the “average” face. Perhaps domain specific expertise is driven by a very well-encoded percept of the categorical average as a baseline so that less individual information needs to be stored for any single item. This would be a very efficient method of storing information and this model would explain findings pertaining to the advantage in the encoding of faces of one’s own ethnicity (Walker & Tanaka, 2003) due to their similarity with a stored average identity.

While this explanation fits the present findings and previous literature, more research is needed in order to establish this as a valid interpretation. Results from the present thesis and previous research are consistent with the idea that not all face features surpass storage limitation, only identity does. However, more face-related features such as age, sex, emotionality, and colour (to name a few) should be investigated alongside identity extraction to discern whether it is identity that is the face-specific exception to the visual system’s capacity limitation. Furthermore, more research is needed to grasp just how the visual system shares information between summary and single face identities, specifically how and if an average face is implicated when reconstructing the percept of a single face. Finally, understanding domain specific expertise as a model of a well-refined perceptual average so that less exemplar-specific information needs to be stored is an interesting hypothesis and warrants further investigation.

The correlational data from Table 1 found no relation between the processing of average and single identity, but there was some evidence (based on a marginally significant trend) that
averaging identity overlapped with averaging viewpoint. These correlational findings are opposite of what is to be expected from the model posed by Haberman, Brady, and Alvarez (2015), stating independent ensemble mechanisms responsible for the processing of high- and low-visual features. Instead, I found a correlation between high-level and low-level visual features (identity and viewpoint), but no correlation within domains (however this is marginally significant and therefore requires more participants). These results may suggest that the independent ensemble mechanisms reside not between the processing of high- and low-level visual features, but between the processing of face-specific and object-specific visual features, which parallels the distinct face and object processing networks for single items (Kanwisher, 2000; Yovel & Kanwisher, 2004; based on double dissociation found in visual object agnosia, e.g. Moscovitch, Winocur, & Behrmann, 1997; and prosopagnosia, e.g. Duchaine & Nakayama, 2005), with rotation (i.e. viewpoint) being treated as a face-specific feature when it applies to faces. This makes sense given that viewpoint extraction in face ensembles may use some part of the face-processing network, seeing that efficiency in reporting a global viewpoint was dependent on whether the mean viewpoint was frontal or not. This ensemble face processing network may also be distinct from a single-face processing network, paralleling the distinct cognitive and neural mechanisms mediating the processing of single objects and object ensemble (Cant & Goodale, 2007; Cant et al., 2015; Cant & Xu, 2017). This would also be consistent with the finding that individuals with prosopagnosia are able to extract average identity from ensemble faces, despite being unable to recognize the identity of a single individual (Yamanashi Leib et al., 2012).

Another major implication for future research is what specifically is meant by an average face identity. Based on face ensemble processing research, there is no doubt that humans can report and recognize an average identity, but the actual subjective percept remains elusive. Does the visual system calculate this average identity implicitly, as an automatic process during passive viewing of faces? Or are reported summary identities based on group familiarity such as family resemblance? Furthermore, the idea of a true average identity may not be as ecologically valid of a concept as other face or object summary percepts. Average emotional expression (Haberman & Whitney, 2007) can give us a crowd’s general disposition towards us, and average viewpoint (Florey et al., 2016) allows us to discern the general direction of attention. Cognitively, it makes sense why certain summary percepts are implicitly extracted from ensemble faces as they
provide a direct benefit in navigating our world. Although summary identity information is extracted, is it represented cognitively as a true feature-based average? What is the utility of a representation of average crowd identity? Future research focusing on summary face extraction should investigate the cognitive percepts of a summary identity, whether a true interpolated average of all features is represented by the visual system, or whether the summary information held is more abstract.

Next, is the linear morphing of features from two anchor faces a valid method to investigate summary identity? Viewpoint, expression, age, and other face-related features are represented continuously and thus true metric averages can be calculated. Does morphing two qualitatively different faces provide a quantitative distribution of face identities, or a series of categorical identities that bear some resemblance? To illustrate this problem, consider a standard colour wheel. If I asked participants to average green (RGB = 0, 255, 0) and orange (RGB = 255, 165, 0) would they give me yellow (RGB = 128, 210, 0)? Or are these still qualitatively different colours based on three quantitative dimensions? This is analogous to face identity, given there are also multiple continuous dimensions making up qualitatively different faces (Nestor et al., 2016). Future research should investigate the nature of qualitative versus quantitative differences across face morphs as it relates to extracting a summary identity. For example they may focus on a single identity dimension (e.g. mouth size) which varies between morph faces, or validate the current standard method where two identities are linearly morphed.

2.7 Limitations

Related to the previous section, one limitation of this study relates to multidimensional extraction versus unidimensional extraction. Various dimensions changed as identities were morphed between anchor faces, but only one dimension changed when viewpoint was manipulated. The large discrepancy in error between viewpoint and identity extraction may reflect differences in the difficulty of extracting multiple dimensions simultaneously for face identity compared with one dimension for viewpoint. These results are the first to directly compare viewpoint and identity extraction from face ensembles. I do not believe, however, that differences in task difficulty confounded the results as participant reports were within the local identity space (their reports were within the first and last local face; illustrated in Figure 1d-e). Also, viewpoint extraction may be more efficient due to being a lower-level feature, but I still cannot rule out
possible floor effects. Future studies which directly compare identity with other face related features are needed to demonstrate whether the difference in error between viewpoint and identity is due to difficulty or a true processing-related difference in the visual system. Considering human observers can distinguish qualitatively different faces so quickly (Richler et al., 2009), it is likely that the difference in error reflects actual processing efficiency and not difficulty related to a floor-effect.

One of the larger limitations concerns the use of computer generated (CG) faces and not real-faces. The use of CG faces confers greater control over modifiable features such as viewpoint and identity and ensures noise that would be present in say adjusting a person or camera viewpoint is absent. It also allows greater control over identity features such as blemishes, symmetry, shape, or other diagnostic features that could bias participant’s judgement of one face over another. However, recent evidence suggests that holistic processing of CG faces suffers in comparison to real faces (Crookes et al., 2015). Specifically, both recognition memory and the own-race effect are diminished for CG faces compared with real faces. The authors give several explanations as to why this may be the case. The first is that there may be less discriminating information in CG faces. The second is that CG face features are unfamiliar, like a different race, due to their difference from real faces. Our visual system is not attuned to processing them as well – much like how it is not attuned to processing faces of other races. Third, the artificiality of the faces evoke somewhat of an uncanny valley feeling. However, it is to be noted that these findings do not invalidate the results present in this thesis (and the field in general). Participants did show discriminability of the face stimuli used. The attenuation of holistic face processing suggested by Crookes and colleagues does not mean there is an absence of it.

The loss of perceptual discriminability in CG faces seems to be a possible reason for the decrease in holistic processing. CG faces, due to absence of fine-grain detail and near perfect symmetry may also be represented as too average. In other words, real-life faces are much more distant from a perceptual average than CG faces are. Still, FFA activation has been observed with FaceGen faces, with an increase in activation as CG faces were more distant from an average face (Said, Dotsch, & Todorov, 2010). FaceGen stimuli have been used in imaging research and have shown appropriate functional encoding of identity in FFA (Sekunova, Fox, Iaria, & Barton, 2008; Xu, Yue, Lescroart, Biederman, & Kim, 2009; Arcurio, Gold, & James, 2012).
The results present here suggest that face-specific processing does take place with CG FaceGen stimuli, even though the strength of the effect may be diminished. As such, replications using real-world faces may actually show stronger effects than what was seen here, but this needs to be investigated empirically. Confounds of using CG stimuli need to be weighed with the benefits of having highly standardized faces, which I believe in this study outweighed potential losses. It is also possible that the fine-grain detail inherent in individual faces is not required as much for the processing of face ensembles, meaning CG faces do not confer a cognitive processing detriment.

Finally, an additional limitation concerns small sample size in experiments here. This has limited power in the validity of whether some results should be treated as a potentially true result despite marginal significance. However, this does not necessarily nullify the validity or reliability of the reported results, as when significant results were reported they were often accompanied by adequate effect sizes.

2.8 Conclusion

My findings show that human observers extract summary identities and viewpoints from face ensembles, and that viewpoint changes interfere with the efficiency of identity extraction efficiency. Changes in local viewpoints also affect how participants extract a summary versus a single identity from face ensembles. Viewpoint extraction may also implicate a face processing network when the viewpoint is of a face, as seen here. It would be interesting to see how patients with prosopagnosia report viewpoint (both single and average) as the majority of research tends to focus on identity related features.

Surprisingly, there was no average-single dissociation for identity processing in Experiment 1, where local viewpoints were held constant when reporting identity. There may be something special about faces, where VWM limitations do not necessarily apply, or allow a much larger capacity of individual faces to be stored in working memory. Another odd finding was that in Experiment 2, there was a dissociation between average and single face identity extraction. The potential interaction of viewpoint on identity processing leads to a host of interesting questions for future research. For example, is face viewpoint extracted by the face processing network? Do high and low level features utilize distinct ensemble coding mechanisms or do face and non-face features utilize distinct ensemble coding mechanisms?
I also find that local deviations in viewpoint lead to better extraction of average identity, likely due to enhanced 3D sampling taking place, but this does not confer an advantage for the extraction of single faces. The speed of processing also seems to be faster for frontal global faces, indicate quicker response onset. Taking longer to complete a response for globally frontal faces may indicate longer time related to fine-tuning the response due to better representation of a summary face across a frontal viewpoint. This is supported by my finding that error for reporting an average identity was less when the mean viewpoint was frontal.

Finally, these results suggest that there may two ensemble systems which are not independent across high versus low level features, but instead show independence for the processing of face and non-face features.

Overall, the present results shed light on developing an understanding of the cognitive mechanisms underlying face ensemble processing. From here, there are numerous extensions, both applied and experimental. Clinical extensions can apply to prosopagnosia, autism, or other conditions where impaired face processing occurs, investigating how ensemble processing systems are impacted. Research in industry can learn how to apply perceptual tricks when rendering crowds in CG (e.g. virtual reality, video game design) where highly detailed features of individual faces may not be needed, saving computational resources. Theoretical work can also continue building on understanding these mechanisms, comparing them to single object and face counterparts in hopes to elucidate whether distinct mechanisms exist for face and object ensembles as they do for single faces and objects. While developmental research exists which suggest a prewired yet immature face processing network, future research could also investigate whether ensemble processing also occurs at some rudimentary level in infants.
<table>
<thead>
<tr>
<th>Task / Feature</th>
<th>Average / Identity</th>
<th>Single / Identity</th>
<th>Average / Viewpoint</th>
<th>Single / Viewpoint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average / Identity</td>
<td>–</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single / Identity</td>
<td>-.098 (.708)</td>
<td>–</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average / Viewpoint</td>
<td>.463 (.061)</td>
<td>-.012 (.953)</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Single / Viewpoint</td>
<td>.046 (.861)</td>
<td>.053 (.840)</td>
<td>.302 (.238)</td>
<td>–</td>
</tr>
</tbody>
</table>
Figure 1.

An illustration of identity space, morphs, and viewpoints. (A) Circular clock-like arrangement of the 75 identities with relative placement of anchor identities. (B) Three sample morphs showing weighted transition between two example anchors. (C) Sample viewpoints for Anchor 1. (D & E) An abstract illustration of the local visual features and their distance to the mean are shown relative to how much of the total feature space they take up. Numbered 1 to 6 are local face features to show their distance to each other, the mean, and the rest of feature space. For reported features, a probe starting feature was always in peripheral feature space. For unreported features, a probe starting feature was always the mean of that feature.
Figure 2.

Sequence of phases in each trial: (A) 400 ms display phase, (B) 500 ms task indication, and (C) report phase (identity or viewpoint). Examples of varying viewpoint with homogenous local identity (A1), and varying identity with homogenous viewpoint (A2) are shown. Next, the cross changes colour to indicate reporting viewpoint (B1) or identity (B2). If participants are tasked to report the average feature of the set (identity or viewpoint), the center probe appears alone (C2). If they are tasked with reporting a single feature (identity or viewpoint), a white oval appears at the location of a previously viewed face, prompting the participant to report the feature of that single face (C1).
Figure 3.

Comparison of participant error for preset mean identities (A) and viewpoints (B). Participants reported mean identity equally well for all of the preset mean identities. In contrast, participants had a trend for higher error when reporting both mean and single viewpoints that derived further away from a frontal viewpoint (regardless of the identity). Error bars represent standard error of the mean.
Figure 4.

Comparison of identity and viewpoint features for average and single extraction tasks. (A) Results show a significant main effect of feature, a significant main effect of task, and a significant interaction. Post hoc analyses revealed a significant difference of reporting single versus average viewpoint, but not for single versus average identity. (B) Results show that participants were faster at initiating their response (RO) to viewpoint, and faster at completing the response (ST) for viewpoint. Error bars represent standard error of the mean.

** p < .01
Figure 5.

Error (MPD) for single feature extraction when compared to the true indicated face versus when compared to the average of the set. This is to test if participants’ selection of a single identity or viewpoint is influenced by the mean of the respective feature. The average of the set biased participants’ reports of single features (both identity and viewpoint) more than the actual single feature they were meant to report. Error bars represent the standard error of the mean.

* $p < .05$

** $p < .01$
Data shows participant’s average error for every preset mean identity in Experiment 2a. Identities are organized based on which anchor set they are between, and by the relative closeness to either anchor. Statistical analyses found that error for morphs within Anchors 2-3 was lower than the other two anchor sets. Error bars represent the standard error of the mean.

* $p < .05$

** $p < .01$
Figure 7.

Average face identity (Experiment 2a, blue line) and single face identity (Experiment 2b, red line) extraction across varying global viewpoints. Curve fitting revealed a significant quadratic function for extraction of average identity, but not single identity, over the varying viewpoints. Specifically, extraction of summary identity was less efficient across progressively non-frontal global viewpoints, but this effect was not seen when extracting a single identity.

* $p < .05$
When averaging identity (A) participants had faster RO when global face viewpoint was frontal and faster ST when global face viewpoint was non-frontal. When reporting a single identity (B) there was no effect of viewpoint on the RO or ST of participant’s total TT.
Figure 9.

Experiment 2b had participants report a single face identity from the set. Here, I compared whether their estimate of the single identity was closer to the true single identity or the average identity across all viewpoints. Analyses revealed non-significant main effect of both Reference (compared to the average or true single identity) and Viewpoint. Error bars represent the standard error of the mean.
References


