# Exploring human visual system: study to aid the development of automatic facial expression recognition framework

Rizwan Ahmed Khan[1,2], Alexandre Meyer[1,2], Hubert Konik[1,3], Saida Bouakaz[1,2]
[1]Université de Lyon, CNRS,
[2]Université Lyon 1, LIRIS, UMR5205, F-69622, France
[3]Université Jean Monnet, Laboratoire Hubert Curien, UMR5516, 42000 Saint-Etienne, France
{Rizwan-Ahmed.Khan, Alexandre.Meyer, Saida.Bouakaz}@liris.cnrs.fr

## Abstract

*This paper focus on understanding human visual system when it decodes or recognizes facial expressions. Results presented can be exploited by the computer vision research community for the development of robust descriptor based on human visual system for facial expressions recognition. We have conducted psycho-visual experimental study to find which facial region is perceptually more attractive or salient for a particular expression. Eye movements of 15 observers were recorded with an eye-tracker in free viewing conditions as they watch a collection of 54 videos selected from Cohn-Kanade facial expression database, showing six universal facial expressions. The results of the study shows that for some facial expressions only one facial region is perceptually more attractive than others. Other cases shows the attractiveness of two to three facial regions. This paper also proposes a novel framework for automatic recognition of expressions which is based on psycho-visual study.*

## 1. Introduction

Communication in any form i.e. verbal or non-verbal is vital to complete various daily routine tasks and plays a significant role in life. Facial expression is the most effective form of non-verbal communication and it provides a clue about emotional state, mindset and intention [10, 9, 6].

Humans have the amazing ability to decode facial expressions across different cultures, in diverse conditions and in a very short time. Human visual system (HVS) has limited neural resources but still can analyze complex scenes in real-time. An explanation for such performance it has been proposed that only some visual inputs are selected by considering "salient regions" [24], where "salient" means most noticeable or most important.

Recently different methods for automatic facial expression recognition have been proposed [15, 23, 14, 22, 20] but none of them try to mimic human visual system in recognizing them. Rather all of the methods, spend computational time on whole face image or divides the facial image based on some mathematical or geometrical heuristic for features extraction. We argue that the task of expression analysis and recognition could be done in more conducive manner, if only some regions are selected for further processing (i.e. salient regions) as it happens in human visual system (HVS).

To determine which facial region(s) is salient according to human vision, we have conducted a psycho-visual experiment. The experiment has been conducted with the help of an eye-tracking system which records the fixations and saccades. It is known that eye gathers most of the information during the fixations [19] as eye fixations describe the way in which visual attention is directed towards silent regions in a given stimuli. We propose novel framework based on HVS for automatic facial expression recognition (FER). Proposed framework creates a new feature space by computationally processing only salient facial regions with Pyramid Histogram of Orientation Gradients (PHOG) [2].

This paper presents result from the experimental study conducted on six universal facial expressions [8]. These six expressions are anger, disgust, fear, happiness, sadness and surprise. These expressions are known as universal facial expressions as these are proved to be consistent across cultures. The aim of this paper is to help computer vision community in the development of robust algorithms that can recognize these facial expression in real time by identifying the potential facial regions that contains discriminative information according to human visual system. As for automatic facial expression recognition, feature selection along with the regions from where these features are to be extracted is one of the most important step.

Thus, our contribution in this study is three fold:

a. We have statistically determined which facial region(s) is salient according to human vision for six universal

facial expressions by conducting a psycho-visual experiment.

b. We have validated the classical results from the domain of human psychology, cognition and perception by a novel approach which incorporates eye-tracker in the experimental methodology protocol. At the same time results have been extended to include all the six universal facial expressions which was not the case in the classical studies.

c. We show that high facial expression recognition (FER) accuracy is achievable by only processing perceptual salient regions (see Section 4 ).

Rest of the paper is organized as follows: all the details related to psycho-visual experiment is described in the next section. Results and analysis of the data gathered from the psycho-visual experiment is presented in the Section 3. Section 4 presents the results of an algorithm that processes only salient facial regions for automatic recognition of expressions. This is followed by the conclusion.

## 2. Psycho-visual experiment

The aim of the experiment was to record the eye movement data of human observers in free viewing conditions. Data was analyzed in order to find which component of the face is salient for specific displayed expression.

### 2.1. Methods

Eye movements of human observers were recorded as subjects watched a collection of 54 videos. Then saccades, blinks and fixations were segmented from each subject's recording.

#### 2.1.1 Participants

Fifteen observers volunteered for the experiment. They include both male and female aging from 20 to 45 years with normal or corrected to normal vision. All the observers were naïve to the purpose of the experiment. They were given only a short briefing about the apparatus and about the expected time required to complete the experiment.

#### 2.1.2 Apparatus and Stimuli

A video based eye-tracker (Eyelink II system from SR Research) was used to record eye movements. The system consists of three miniature infrared cameras with one mounted on a light-weight headband for head motion compensation and the other two mounted on arms attached to headband for tracking both eyes. Stimuli were presented on a 19 inch CRT monitor with a resolution of 1024 x 768 and a refresh rate of 85 Hz. A viewing distance of 70cm was maintained resulting in a 29° x 22° usable field of view as done by Jost et al. [13]. For the experiment, videos from the extended Cohn-Kanade (CK+) database [16] were selected. CK+ database contains 593 sequences across 123 subjects which are FACS [11] coded at the peak frame. Database consists of subjects aged from 18 to 50 years old, of which 69% were female, 81% Euro-American, 13% Afro-American and 6% others. Each video (without sound) showed a neutral face at the beginning and then gradually developed into one of the six universal facial expression. Fifty four videos were selected for the experiment, with the criteria that videos should show both male and female actors and posed facial expression should not look unnatural. Another consideration while selecting the videos was to avoid such sequences where the date/time stamp is not recorded over the chin of the subject [18].

### 2.2. Procedure

The experiment was performed in a dark room with no visible object in observer's field of view except a stimulus. Stimuli were presented in random order. It was carefully monitored that an experimental session should not exceed 20 minutes, including the calibration stage. This was taken care of in order to prevent participant's lose of interest or disengagement over time.

#### 2.2.1 Eye movement recording

Eye position was tracked at 500 Hz with an average noise less than 0.01°. Head mounted eye-tracker allows flexibility to perform experiment in free viewing conditions as the system is designed to compensate for small head movements. Then the recorded data is not affected by head motions and participants can observe stimuli with no severe restrictions. Severe restrictions in head movements have been shown to alter eye movements and can lead to noisy data acquisition and corrupted results [4].

## 3. Results and Discussion

### 3.1. Gaze map construction

The most intuitively revealing output that can be obtained from the recorded fixations data is to obtain gaze maps. For every frame of the video and each subject $i$, the eye movement recordings yielded an eye trajectory $T^i$ composed of the coordinates of the successive fixations $f_k$, expressed as image coordinates $(x_k, y_k)$:

$$T^i = (f_1^i, f_2^i, f_3^i, ....) \tag{1}$$

As a representation of the set of all fixations $f_k^i$, a human gaze map $H(\mathbf{x})$ was computed, under the assumption that this map is an integral of weighted point spread functions $h(\mathbf{x})$ located at the positions of the successive fixations. It


Figure 1. Examples of gaze maps for six universal expressions. Each video sequence is divided in three mutually exclusive time periods. First, second and third columns show average gaze maps for the first, second and third time periods of a particular stimuli respectively.

is assumed that each fixation gives rise to a normally (Gaussian) distributed activity. The width $\sigma$ of the activity patch was chosen to approximate the size of the fovea. Formally, $H(\mathbf{x})$ is computed according to Equation 2:

$$H(\mathbf{x}) = H(x, y) = \sum_{i=1}^{N_{subj}} \sum_{f_k \in T^i} \exp\left(\frac{(x_k - x)^2 + (y_k - y)^2}{\sigma^2}\right) \tag{2}$$

where $(x_k, y_k)$ are the spatial coordinates of fixation $f_k$, in image coordinates. In Fig. 1 gaze maps are presented as the heat maps where the colored blobs / human fixations are superimposed on the frame of a video to show the areas where observers gazed. The longer the gazing time is, the warmer the color is.

As the stimuli used for the experiment is dynamic i.e. video sequences, it would have been incorrect to average all the fixations recorded during trial time (run length of video) to construct gaze maps as this could lead to biased analysis of the data. To meaningfully observe and analyze the gaze trend across one video sequence we have divided each video sequence in three mutually exclusive time periods. The first time period correspond to initial frames of the video sequence where the actor's face has no expression i.e. neutral face. The last time period encapsulates the frames where the actor is showing expression with full intensity (apex frames). The second time period is a encapsulation of the frames which has a transition of facial expression i.e. transition from the neutral face to the beginning of the desired expression (i.e neutral to onset of the expression). Then the fixations recorded for a particular time period are averaged across 15 observers.

### 3.2. Observations from the gaze maps

Fig. 1 gives the first intuition that gazes across all the observers are mostly attracted towards three facial regions i.e. eyes, nose and mouth for all the six universal facial expressions.

Secondly, gaze maps presented in the Fig. 1 suggests the saliency of mouth region for the expressions of happiness and surprise. It can be observed from the figure that as the two said expressions becomes prominent(second and third time periods) most of the gazes are attracted towards only one facial region and that is the mouth region. The same observation can be made for the facial expressions of sadness and fear but with some doubts. For the expressions of anger and disgust it seems from the gaze maps that no single facial region emerged as salient, as the gazes are attracted towards two to three facial regions even when the expression was show at its peak.

### 3.3. Substantiating observations through statistical analysis

In order to statistically confirm the intuition gained from the gaze maps about the saliency of different facial region(s) for the six universal facial expressions we have calculated the average percentage of trial time observers have fixated their gazes at specific region(s) in a particular time period (definition of time period is same as described previously). The resulting data is plotted in Fig. 2.

Fig. 2 confirms the intuition that the region of mouth is the salient region for the facial expressions of happiness and surprise. Third time period in the figure corresponds to the time in the video when the expression was shown at its peak. It can be easily observed from the figure that

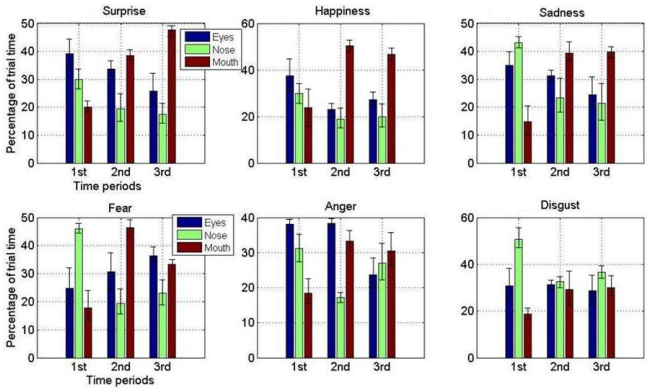


Figure 2. Time period wise average percentage of trial time observers have spent on gazing different facial regions. The error bars represent the standard error (SE) of the mean. First time period: initial frames of video sequence. Third time period: apex frames. Second time period: frames which has a transition from neutral face to particular expression.

as the expression of happiness and surprise becomes more prominent, the humans tend to fixate their gazes mostly on the facial region of mouth. This result is consistent with the results by Cunningham et al. [5], Nusseck et al. [17] and Boucher et al. [3].

Cunningham et al. [5] and Nusseck et al. [17] have reported that the recognition of facial expression of sadness requires a complicated interaction of mouth and eye region along with rigid head motion, but the data we have recorded from experiment and plotted in Fig. 2 shows that human visual system tends to divert its attention towards the region of mouth as the expression becomes prominent. The fact can be observed in the second and third time periods. But still the contribution of eye and nose regions cannot be considered negligible as in terms of percentage for the third time period observers have gazed around 40 percent of the trial time at the mouth region and around 25 percent each at the eye and nose regions. Fig. 3 shows the average gaze maps from the 15 observers for the expression of sadness and it confirms the fact that the facial region of mouth get more attention than the facial regions of eye and nose.

Facial expression of disgust shows quite random behavior. Even when the stimuli was shown at its peak, observers have gazed all the three regions in approximately equal proportions. The only thing that can be point out from the Fig. 2 is that there is more attraction towards the nose region and that could be due to the fact that wrinkles on the nose region becomes prominent and attracts more attention when the stimuli shows maximum disgust expression. Ironically, Cunningham et al.[5] and Nusseck et al.[17] while discussing the results for the expression of disgust have not considered the contribution of the nose region which has came out to be little bit more prominent then the other two facial regions in terms of saliency from the results of the current study. Fig. 4 presents the average gaze maps from the 15 observers for the facial expression of disgust. From the gaze maps of two presented video sequences it is evident that the wrinkles in the nose region gets bit more attention than the other two facial regions.

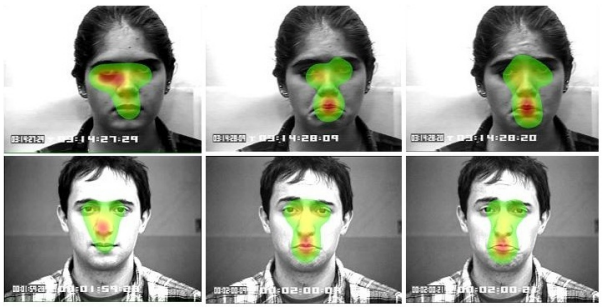

Figure 3. Gaze maps for the facial expression of sadness. First, second and third columns shows average gaze maps for the first, second and third time periods of stimuli respectively. Each row corresponds to different stimuli/video sequence for “sadness”.

For the expression of fear, facial regions of the mouth and eyes attract most of the gazes. From Fig. 2 it can be seen that in the second trial time period (period correspond to the time when observe experiences the change in face presented in stimuli toward the maximum expression) observers mostly gazed at the mouth region and in the final trial period eye and mouth regions attracts most of the attention. Hanawalt [12] reported that the expression of fear is mostly specified by the eye region but our study shows the interaction of facial regions of mouth and eyes for the fear. Fig. 5 shows the average gaze maps for the expression of fear and these gaze maps confirms the argument that “fear” is accompanied with the interaction of mouth and eye regions.

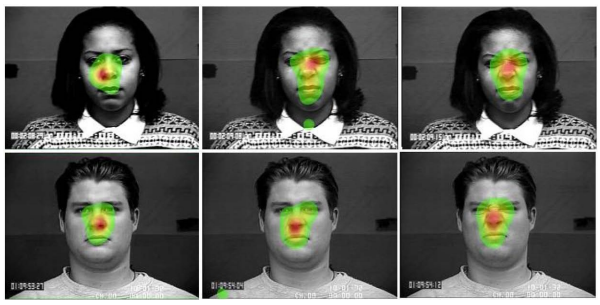

Figure 4. Gaze maps for the facial expression of disgust. First, second and third columns shows average gaze maps for the first, second and third time periods of stimuli respectively. Each row corresponds to different stimuli/video sequence for “disgust”.

Boucher et al. [3] in 1975 wrote that “Anger differs from the other five facial expressions of emotion in being ambiguous” and this observation holds for the current study

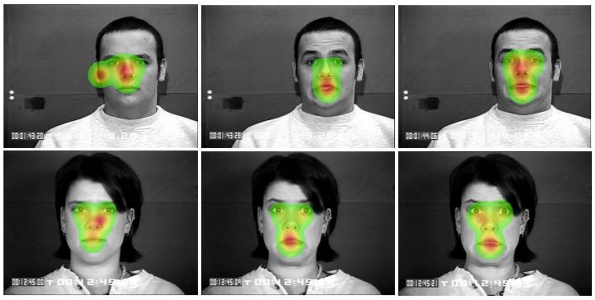

Figure 5. Gaze maps for the facial expression of fear. First, second and third columns shows average gaze maps for the first, second and third time periods of stimuli respectively. Each row corresponds to different stimuli/video sequence for "fear".

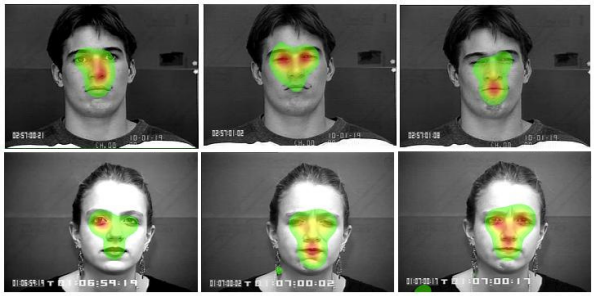

Figure 6. Gaze maps for the facial expression of anger. First, second and third columns shows average gaze maps for the first, second and third time periods of stimuli respectively. Each row corresponds to different stimuli/video sequence for "anger".

as well. "Anger" shows complex interaction of eye, mouth and nose regions with out any specific trend. This fact is evident from the Fig. 6 as observers have gazed at different regions of the face for the two stimuli showing "anger". But one thing is common for all the stimuli in Fig. 6 that for "anger" no facial region emerges as the salient but all the three regions are gazed interchangeably even when the expression was shown at its peak/apex.

## 4. Automatic Facial Expression Recognition Framework

We conducted an experiment to test an idea that algorithmically expressions can be recognized by processing only perceptual salient regions. The proposed algorithm / framework creates a novel feature space by extracting and concatenating Pyramid Histogram of Orientation Gradients (PHOG) [2] features from the perceptual salient facial regions i.e. mouth and eye regions. PHOG features are selected as they have proven to be highly discriminative for FER task [1, 7].

We conducted facial expression recognition experiment on the two databases: (1) extended Cohn-Kanade (CK+) database [16] (2) MMI facial expression database [18]. The performance of the algorithm was evaluated using four classical classifiers i.e. "Support vector machine (SVM)" with $\chi^2$ kernel and $\gamma$=1, "C4.5 Decision Tree" with reduced-error pruning, "Random Forest" of 10 trees and "2 Nearest Neighbor (2NN)" based on Euclidean distance. The parameters of the classifiers were determined empirically. In both the experiments, frames which covers the status of onset to apex of the expression were used. The same has been done by Yang et al. [22]. Region of interest was obtained automatically by using Viola-Jones object detection algorithm [21] and processed to obtain PHOG feature vector. In both the experiments, average recognition accuracy for the six universal facial expressions [8] was calculated using *10*-fold cross validation.

### 4.1. Results: Automatic Expression Recognition

For the first experiment we used all the 309 sequences from the CK+ database which have FACS coded expression label. The proposed framework achieved average recognition rate of 95.3%, 95.1%, 96.5% and 96.7% for SVM, C4.5 decision tree, random forest and 2NN respectively.

For the second experiment we used 238 clips of 28 subjects from the MMI database (Part II of the database). These clips also includes people wearing glasses. For this database proposed framework achieved average recognition rate of 93.2%, 91.2%, 94.3% and 95.8% for SVM, C4.5 decision tree, random forest and 2NN respectively.

One of the most interesting aspects of our approach is that it gives excellent results for a simple 2NN classifier which is a non-parametric method. This points to the fact that framework do not need computationally expensive methods such as SVM or decision trees to obtain good results. In general, the proposed framework achieved high expression recognition accuracies irrespective of the classifiers, proves the descriptive strength of the features.

## 5. Conclusion

The presented experimental study provides the insight into which facial region(s) emerges as the salient according to human visual attention for the six universal facial expressions. Eye movements of fifteen human observers were recorded using eye-tracker as they watch the stimuli which was taken from the widely used Cohn-Kanade facial expression database. Conclusions drawn from the experimental study are summarized in Table 1. Presented results can be used as the background knowledge by the computer vision community for deriving robust descriptor for the facial expression recognition (FER) as for FER, feature selection along with the regions from where these features are to be extracted is one of the most important step. Secondly, processing only salient regions could help in reducing computational complexity of the FER algorithms.

This paper also presented a framework for automatic and reliable facial expression recognition. Framework is based

Table 1. Summary of the facial regions that emerged as salient for six universal expressions

| Facial expression | Salient facial region(s) |
|---|---|
| Happiness | Mouth region. |
| Surprise | Mouth region. |
| Sadness | Mouth and eye regions. Biased towards mouth region. |
| Disgust | Nose, mouth and eye regions. Wrinkles on the nose region gets little more attention than the other two regions. |
| Fear | Mouth and eye regions. |
| Anger | Mouth, eye and nose regions. |

on a initial study of human vision. With the proposed framework high recognition accuracy is achieved by processing only perceptually salient region of face. Proposed framework can be used for real-time applications since our unoptimized Matlab implementation run at 4fps which is enough as facial expressions does not change abruptly.

## 6. Acknowledgment

This project is supported by the Région Rhône-Alpes, France.

## References

[1] Y. Bai, L. Guo, L. Jin, and Q. Huang. A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition. In *International Conference on Image Processing*, 2009. 5

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *6th ACM International Conference on Image and Video Retrieval*, pages 401–408, 2007. 1, 5

[3] J. D. Boucher and P. Ekman. Facial areas and emotional information. *Journal of communication*, 25:21–29, 1975. 4

[4] H. Collewijin, M. R. Steinman, J. C. Erkelens, Z. Pizlo, and J. Steen. *The Head-Neck Sensory Motor System*. Oxford University Press, 1992. 2

[5] D. W. Cunningham, M. Kleiner, C. Wallraven, and H. H. Bülthoff. Manipulating video sequences to determine the components of conversational facial expressions. *ACM Transactions on Applied Perception*, 2:251–269, 2005. 4

[6] F. De la Torre and J. F. Cohn. *Guide to Visual Analysis of Humans: Looking at People*, chapter Facial Expression Analysis. Springer, 2011. 1

[7] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using PHOG and LPQ features. In *IEEE Automatic Face and Gesture Recognition Conference FG2011, Workshop on Facial Expression Recognition and Analysis Challenge FERA*, 2011. 5

[8] P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*, pages 207–283. Lincoln University of Nebraska Press, 1971. 1, 5

[9] P. Ekman. Facial expression of emotion. *Psychologist*, 48:384–392, 1993. 1

[10] P. Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. W. W. Norton & Company, New York, 3rd edition, 2001. 1

[11] P. Ekman and W. Friesen. The facial action coding system: A technique for the measurement of facial movements. *Consulting Psychologist*, 1978. 2

[12] N. Hanawalt. The role of the upper and lower parts of the face as the basis for judging facial expressions: Ii. in posed expressions and "candid camera" pictures. *Journal of General Psychology*, 31:23–36, 1944. 4

[13] T. Jost, N. Ouerhani, R. Wartburg, R. Müri, and H. Hügli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding. Special Issue on Attention and Performance in Computer Vision*, 100:107–123, 2005. 2

[14] I. Kotsia, S. Zafeiriou, and I. Pitas. Texture and shape information fusion for facial expression and facial action unit recognition. *Pattern Recognition*, 41:833–851, 2008. 1

[15] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24:615–625, 2006. 1

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kande dataset (ck+): A complete facial expression dataset for action unit and emotion-specified expression. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2010. 2, 5

[17] M. Nusseck, D. W. Cunningham, C. Wallraven, and H. H. Bülthoff. The contribution of different facial regions to the recognition of conversational expressions. *Journal of vision*, 8:1–23, 2008. 4

[18] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo*, 2005. 2, 5

[19] U. Rajashekar, L. K. Cormack, and A. Bovik. Visual search: Structure from noise. In *Eye Tracking Research & Applications Symposium*, pages 119–123, 2002. 1

[20] M. Valstar, I. Patras, and M. Pantic. Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 76–84, 2005. 1

[21] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001. 5

[22] P. Yang, Q. Liu, and D. N. Metaxas. Exploring facial expressions with compositional features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1, 5

[23] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 29:915–928, 2007. 1

[24] L. Zhaoping. Theoretical understanding of the early visual processes by data compression and data selection. *Network: computation in neural systems*, 17:301–334, 2006. 1