Hierarchical Combination of Object Models using Mutual Information

Hannes Kruppa and Bernt Schiele
Perceptual Computing and Computer Vision Group
ETH Zurich, Switzerland
{kruppa,schiele}@inf.ethz.ch
http://www.vision.ethz.ch/pccv

Abstract

Combining different and complementary object models promises to increase the robustness and generality of today's computer vision algorithms. This paper introduces a new method for combining different object models by determining a configuration of the models which maximizes their mutual information. The combination scheme consequently creates a unified hypothesis from multiple object models "on the fly" without prior training. To validate the effectiveness of the proposed method, the approach is applied to the detection of faces combining the output of three different models.

1 Introduction

Any object model has its own strengths and weaknesses depending on the context and the dynamics of the environment. This becomes evident in virtually any model-based computer vision task such as object detection, tracking or recognition. As for recognition there exist a considerable number of highly successful algorithms. Each of the approaches however has its own inherent limitations and weaknesses. Similar observations hold for tracking and detection of objects. In this paper we argue that in order to overcome such inherent limitations one has to integrate and combine different, complementary approaches and object models. While there are many computer vision algorithms for computing various kinds of object models, work on the *combination* of different models is still in its infancy. Since no single model is robust and general enough to cover all possible environmental conditions, their combination promises to increase robustness and generality.

The ultimate goal of the proposed approach is to overcome the limitations of individual models by combining multiple models at different levels. In order to integrate and combine complementary object models, this paper proposes a general framework based on mutual information which is used to measure mutual agreement between different object models. Rather than learning a static integration scheme, the algorithm determines model configurations which maximize agreement among the employed models and the current image data. Since the image data is used directly, the integration mechanism dynamically adapts to environmental changes for example in lighting conditions. The framework therefore allows to combine different models dynamically and 'on-the-fly'. This makes the approach general and easily extendible to new object models.

BMVC 2001 doi:10.5244/C.15.12

The main focus of this paper is the important and interesting question of how to combine different object models effectively when facing a dynamic, real-world environment. The remainder of this paper is organized as follows: After a short review of related work section 2 introduces the general framework for model combination based on mutual information. Section 3 describes a case study in which human faces are detected by combining three distinct face models. In section 4 the robustness of the combination method is demonstrated. We also give an example of how the algorithm selects a skin color model which is most appropriate for the current environmental conditions.

1.1 Related Work

Probably the most common approach to integrate information is to accumulate all responses (raw data or features) into a single vector. Taking a number of these vectors, standard machine learning techniques [4] can be used to learn a joint probabilistic model. Bayesian approaches can be taken to statistically model dependencies between the different data sources and sensors. However, the amount of training data required may be prohibitive [2]. [1] proposes a hierarchical mixture of experts [7] in which mixtures of appearance models are learned for the classification of objects. The common drawback of these approaches is that the integration is static in the sense that we cannot change the weighting of responses dynamically depending, for example, on their usefulness or the environmental conditions.

Combining different classifiers is a standard problem in pattern recognition (see for example [15, 8, 6]). Typically different classifiers are trained individually and their parameters are fixed thereafter. It is the combiner's task to learn and choose the appropriate combination mechanisms depending on particular situations. In that sense only the combiner itself may be able to increase the robustness and generality of the individual classifiers. However, the results will most often be sup-optimal since the combiner does not have access to the classifiers themselves and cannot adapt their parameters. As before the combination scheme is fixed and static during runtime.

2 Hierarchical Model Combination Using Mutual Information

This paper proposes a general method for the combination of multiple object models using the information-theoretic principle of *mutual information*. Mutual information has been used previously in computer vision, as for example in medical image registration [14], for selection of most discriminant viewpoints of objects and in audio-visual speech acquisition [11]. As detailed below, mutual information can be used to measure the *mutual agreement* between two object models. In order to combine multiple models a hierarchy of pairwise model combinations is used.

The mutual information of two random variables X and Y with a joint probability mass function $p(x, y)$ and marginal probability mass functions $p(x)$ and $p(y)$ is defined as [3]:

$$I(X;Y) = \sum_{x_i, y_j} p(x_i, y_j) log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{1}$$

Here, the probabilities in expression 1 can be directly derived from a pair of distinct visual object models. To undermine the relevance of mutual information in the context of object model combination, we briefly refer to the well-known Kullback-Leibler divergence. The KL-divergence between a probability mass function $p(x, y)$ and a distinct probability mass function $q(x, y)$ is defined as:

$$D(p(x,y)||q(x,y)) = \sum_{x_i,y_j} p(x_i, y_j) log \frac{p(x_i, y_j)}{q(x_i, y_j)} \tag{2}$$

Although the Kullback-Leibler divergence (also called relative entropy or information divergence) is not symmetric and does not satisfy the triangle inequality, it is often useful to think of it as a "distance" between distributions [3]. By defining $q(x, y) = p(x)p(y)$ the mutual information can be written as the KL-divergence between $p(x, y)$ and $p(x)p(y)$:

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \tag{3}$$

Mutual information therefore measures the "distance" between the joint probability $p(x, y)$ and the probability $q(x, y) = p(x)p(y)$, which is the joint probability under the assumption of independence. Conversely, it measures mutual dependency or the amount of information one object model contains about another. As a result mutual information can be used to measure *mutual agreement* between object models.

In the following we assume that for each subregion of the input image, each model determines the probability that the object of interest is either present or absent. This representation is very general and can be satisfied by nearly any object model. $p(x)$ is calculated based on the first object model and covers two cases, namely the presence of the object $p(x_0)$ or its absence $p(x_1) = 1 - p(x_0)$, respectively. The probability $p(y)$ is derived from the second object model analogously, also with the two described cases. Finally, for the joint probability $p(x, y)$ both models and all four cases are taken into account.

Typically, the object of interest can be associated with a characteristic parameter range of an object model. For example, in the case of a color model, the parameter range may be given by a particular subspace of the total color-space. Note that each parameter configuration results in distinct probabilities $p(x)$ and $p(y)$ and consequently, in a distinct mutual information value. Therefore, one can determine a configuration $(\alpha^\star, \beta^\star)$ which maximizes mutual agreement between the employed models and the input data by maximizing the mutual information over the object-specific joint parameter space:

$$(\alpha^\star, \beta^\star) = \arg I_{max}(X;Y) = \arg \max_{\alpha,\beta} I(X;Y) \tag{4}$$

with α and β describing the object-specific parameter space of a model pair. For example the parameters of the facial shape model in the experiments described below are the size and the location of the face within the image. By maximizing mutual information with a second, complementary face model the algorithm detects and locates faces in the image.

At each stage of a hierarchy, the algorithm computes a ranking of parameter configurations which maximize mutual information. This ranking is then used as input for the next stage in the hierarchy where mutual information can be used again to find the best combined parameter configurations. The resulting algorithm is modular and can be easily extended to new object models.

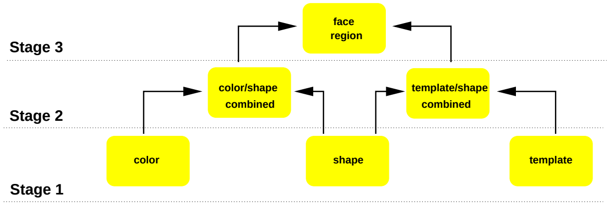

Figure 1: Multi-stage usage of maximization of mutual information for combining object models

The hierarchical concept is depicted in figure 1 which shows the architecture used in the case study described below. In this case study, the following three object models are combined pairwise in order to detect human faces: a skin color model, a shape model and a template matcher. In stage one the probability maps are calculated based on the color model and the template matcher. Stage two combines the color model with the facial shape model by maximizing mutual information. The template matcher and the facial shape model are also combined in stage two. Finally stage three combines both results again by maximizing mutual information.

Obviously other groupings would be meaningful as well. The proposed grouping however ensures that the combined hypotheses on stage two can be represented as a single condensed region of probabilities. This will be further explained in the next sections. Also, it would be possible to combine all models in a single maximization step. However, using pairwise combinations enables the definition of separate and independent parameter constraints for each pair which reduces the size of the joint parameter space and therefore speeds up sampling.

3 Case Study: Experiments on Face Detection

Face detection is a well studied problem in computer vision. The diversity of approaches ranges from neural networks [10], over support vector machines, use of facial features and geometric constraints, density estimation to clustering. These and other approaches have reached impressive detection rates depending on the respective context. Despite these considerable successes in face detection one can identify specific strengths and weaknesses of the underlying models. For example the approach of Rowley et al. [10] only detects frontal upright faces. Schneiderman and Kanade [12] introduce a different approach based on wavelet coefficients which can handle profile views. They explicitly note that particular models (i.e. use of eigenvectors) seem to be more appropriate for frontal faces whereas their approach based on wavelet coefficients is more successful on profiles.

Rather than to refine any of the above approaches we argue that one should focus on the combination of the most successful and promising approaches in order to overcome their inherent limitations. As stated in the introduction, this paper therefore focuses on the combination of different models. In the experiments described below the limitations of the facial models used are widely understood. In this respect, they are well-suited to highlight the utility of the proposed integration scheme.

3.1 Three models for face detection

Three simple models for face detection are used in the experiments described below: a skin color model, a face template matcher and a shape model. Note that taken together, the three models embody desirable and complementary properties for face detection. In the experiments individual models are challenged in order to analyze the robustness of the proposed combination mechanism.

The first model is a Gaussian skin color model (see for example [9]). More specifically the color model is represented by a single Gaussian of the hue, saturation and intensity values with diagonal covariance matrix. The model's parameters are trained on several images using the maximum likelihood principle. Several such skin color models are trained for different environments such as to cover various indoor and outdoor illumination conditions. Each of these models allow to calculate probabilities of skin color for every pixel in an image.

The second model is based on template matching. A face template is correlated with all local subregions of the image having the expected size of the face. The current implementation uses several templates of frontal faces from different people. The correlation results are normalized and projected onto probabilities using a sigmoid function. Thus, the probability maps generated for each template represent the probability distribution of face centers in the image.

As a third model, facial shape is used. Essentially, this model imposes the constraint that the face region be compact and elliptical. The parameters are the position of the face region (given in pixel coordinates) and the two main axes of the ellipse. The model could be extended by a fifth parameter, namely the rotation of the ellipse which has been omitted in the following. The third model is implemented as an ellipse with high probabilities in the center and decreasing probabilities towards the edges of the ellipse. Probabilities outside the elliptical region are set to zero.

3.2 Illustration

This section describes the combination of the three face models in more detail illustrated with an example. Figure 2 shows the original image with two people and the different stages of the algorithm from left to right. As can be seen in figure 1, the combination has three stages. In stage one the skin color model as well as the template matcher are applied to the input image depicted on the left. The resulting probability maps are shown with white indicating high probability. Face color probabilities are depicted on top, the template matcher's probabilities are shown underneath.

In the second stage, the skin color model is combined with facial shape and the template matcher is combined with facial shape in parallel. That is, the mutual information between the facial shape model and either probability map is maximized. To this end, the joint configuration spaces of both model pairs are sampled. More specifically, the algorithm samples over all possible face shape configurations and locations in the probability maps. In the current implementation, image locations are examined according to their probability. The first 100 resulting maxima are visualized in figure 2 by drawing

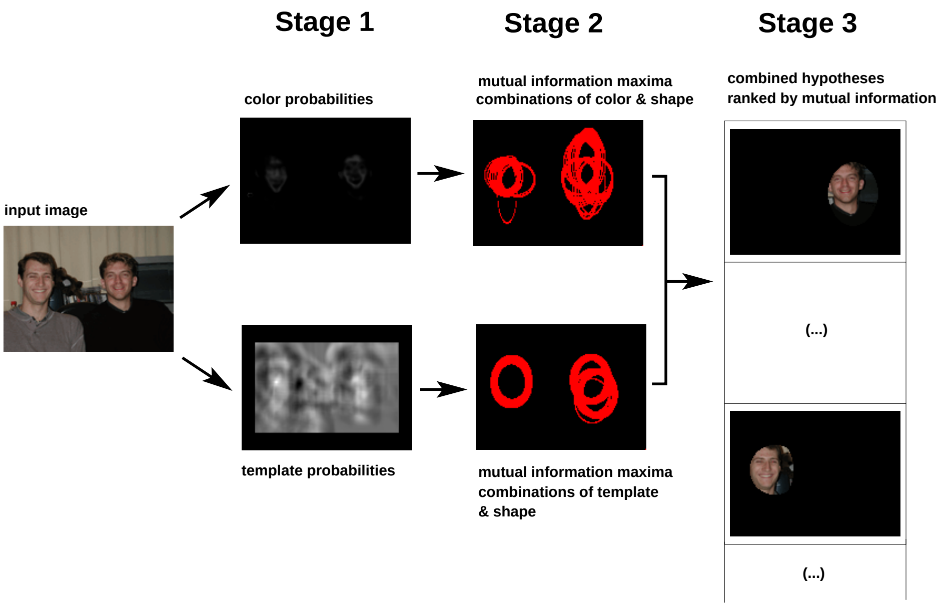
Stage 1
Stage 2
Stage 3
input image
color probabilities
template probabilities
mutual information maxima combinations of color & shape
mutual information maxima combinations of template & shape
combined hypotheses ranked by mutual information
(...)
(...)

Figure 2: Combining three object models for face detection and localization

the facial shape at the associated locations. The upper image contains the ellipses from the combination of the skin color model and the face shape model. The image below corresponds to the combination of the template matcher and face shape. In this first rather simple example, the results from both pairwise combinations are in close agreement, as can be seen in the figure.

In the third and final stage maxima from combining color and shape are paired with maxima from combining shape and template. More specifically, the mutual information of all possible pairs of maxima is computed and used to rank each individual pair. The final hypothesis is then visualized by intersecting the two shape models with the highest mutual information. As shown in figure 2 on the right, the face of the right person in the original image has been successfully located by this procedure.

Also the second person is extracted successfully since the algorithm is able detect and locate multiple instances of the object of interest, in this case, multiple faces: closely co-located maxima are merged into a single hypothesis of a detected face, while spatially separated maxima indicate the presence of multiple faces. Here, the next few mutual information maxima are in the same location with similar but slightly different shapes, also containing the face of the right person. Then, the first spatially separated maxima captures the location of the left face in the input image. Note that the actual value of mutual information can be interpreted as a degree of confidence associated with the generated hypotheses.

4 Experimental Results

The section describes examples where the individual face models are challenged in order to test the robustness of the proposed combination scheme. In a first scenario the face color model is challenged by using a camera flash (see figure 3). In a second scenario strong shadows on the faces cause difficulties for the template matcher (figure 4). The results suggest in both cases that the proposed approach is robust enough to handle these failures and weaknesses of individual models.

4.1 Using mutual information to combine face models

The input image of the first scenario is shown on the left of figure 3. The picture is taken with the camera's flash. There is also daylight entering through a window in the background. As can be seen, the flash causes reflections on the window's metal frame. Since the flash strongly illuminates both faces the colors of both faces are distorted. Even though the skin color model is challenged by this lighting condition the algorithm still finds both faces as shown on the right of the figure.

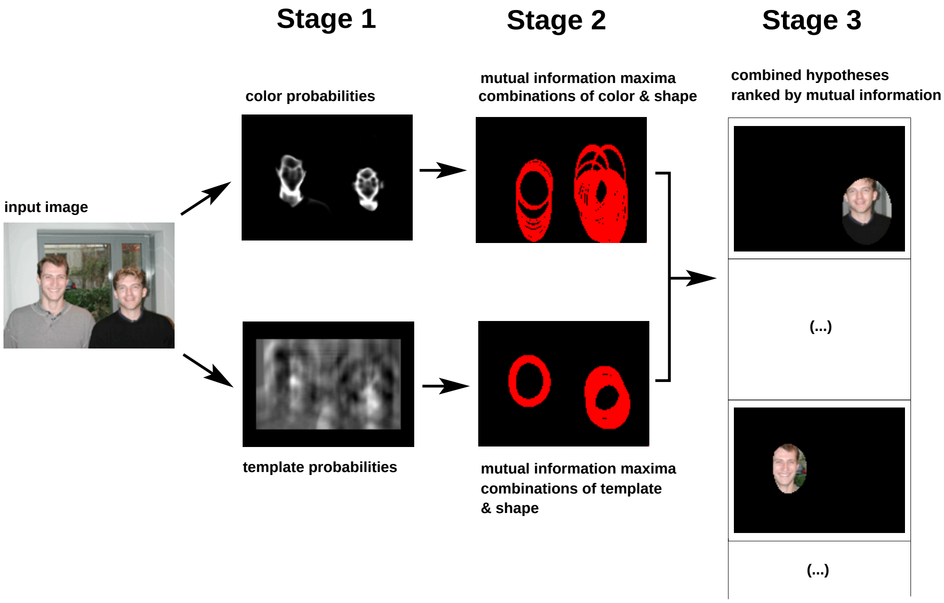

Figure 3: Input image taken with camera flash which challenges the skin color model. The combination with shape and template matcher allows robust detection of both faces

The intermediate stages of the computation as shown in figure 3 are discussed in the following. The skin color probabilities show (stage 1) that the color model can only capture the faces' edges while large regions within the faces have zero probability (false negatives). Note also the occurrence of some false positives along the shoulder of the left person also caused by the flash. The template matcher's probabilities indicate, however, that it is not as much affected by the flashlight: regions of high probability coincide with the actual locations of faces. In the next stage of the algorithm either probability map is combined with the shape model using mutual information. The combination of color and shape results in several compact face regions. This underlines that the shape model essentially injects the desired compactness property, which the color model could not establish because of the flash. In this scenario, the combination of template and shape yields several useful region candidates for the two faces.

The final hypotheses generated by stage three are visualized as described in section 3.2. The hypothesis with the highest mutual information captures the face location of the right person. The face of the second person is contained in the next local maximum of mutual information that is spatially separated from the first. In this experiment the algorithm is capable to deal with the flash light situation robustly and overcomes the inherent limitations of the color model.

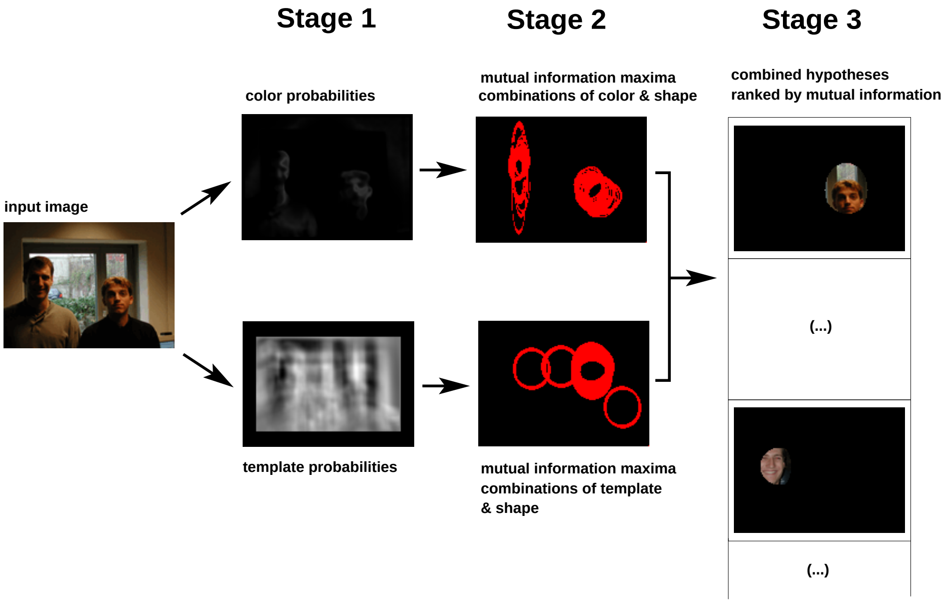

Figure 4: Image with light from the side causing shadows on both faces. Even though the face template matcher fails the two faces are detected by the combination scheme

Figure 4 shows another input image on the left. In this scenario the template matcher is challenged since most of the light comes from the side creating significant shadows on both faces. In this case the template matcher cannot handle the local shifts in intensities caused by the shadows. Hence, the resulting probability map contains a large number of false positives and false negatives. In this situation also the probability map of the skin color model exhibits a large number of false negatives and false positives. Also the shirt of the left person as well as a larger background region have non-zero probabilities.

Given this input, it is not surprising that combining template with shape in the second stage produces many hypotheses in wrong places – but it also yields some, which are correctly located over the actual faces. The combination of color and shape produces many useful intermediate hypotheses, despite many false positives and false negatives. This can be explained by the fact that the probability map of the color model still contains enough information that can be put in agreement with the shape model's compactness.

Note that analogously to the template output, the color probability map is hard to interpret directly, and only after the combination with the shape model the relevant information is available.

As before, the mutual information among intermediate hypotheses is computed to generate and rank the final hypotheses. Since this step eliminates disagreeing shape configurations the information of the color-shape combination is preserved while ill-formed hypotheses from the shape-template combination are discarded. It is also interesting to note that the mutual information value associated with the final hypothesis is lower than in the example before. This suggests that the value of mutual information can be used directly as a measure of confidence for the generated hypothesis, since it is more difficult to achieve agreement among all models in this scenario. As can be seen, the algorithm deals with the shadow situation successfully and is able to detect both faces robustly.

In the above experiments we explicitly challenged each of the employed models in order to show how the proposed combination algorithm can overcome limitations of individual models. Most results reported in the literature however are describing cases where the individual models perform well. It is therefore difficult to compare our results to existing face detection systems. Furthermore a direct quantitative comparison is difficult since employed benchmark face databases consist only of gray-scale images [10, 13]. The emphasis of this paper is therefore the illustration of the novel mutual information-based combination which attains increased robustness by employing multiple object models. To this end, the face detection task serves as a demonstration case: multiple instances are robustly detected even in scenarios which pose significant real-world challenges. As a sidenote, mutual information may also be used for switching between different model instances [5].

5 Conclusion and Future Work

This paper argues that even the most sophisticated *stand-alone* models have specific weaknesses. However, the combination of different complementary models opens the possibility to overcome limitations of individual models and to increase robustness with respect to environmental changes. This paper therefore introduced a hierarchical framework for combining different object models based on maximization of mutual information. Object models are combined into a single unified hypothesis *on-the-fly* without need for specific training. The framework thus allows to easily add, reuse, remove and replace object models in a flexible manner.

In order to validate the proposed approach it has been applied to face detection. In this experimental case study three face models (color, shape and template) are combined on-the-fly in an hierarchical fashion. By deliberately challenging the weaknesses of the individual models the robustness of the actual *combination* of models could be analyzed.

As the value of mutual information can be interpreted as a confidence measure. This paper therefore showed how mutual information can be used to select the best available color model depending on the environmental conditions. In the future we are planning to use this for optimally initializing an object tracker. Mutual information may be used also for the combination of models for recognition.

References

[1] C. Bregler and J. Malik. Learning appearance based models: Mixtures of second moment experts. In *Advances in Neural Information Precessing Systems*, 1996.

[2] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. In *Proceedings of the 2nd International Conference on Audio-Visual Biometric Person Authentication*, 1998.

[3] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc, 1991.

[4] R. Duda and P. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., 1973.

[5] M. Spengler H. Kruppa and B. Schiele. Context-driven model switching for visual tracking. In *9th International Symposium on Intelligent Robotic Systems'01*, July 2001.

[6] A. Jain, P. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, January 2000.

[7] M.I. Jordan and R.A. Jacobs. Hierachical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2), March 1994.

[8] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, March 1998.

[9] J.Coutaz N.Oliver, F.Berard and A. Pentland. Lafter: Lips and face tracker. In *Proceedings IEEE Conf. Computer Vision and Pattern Recognition*, pages 100–110, 1997.

[10] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, 1998.

[11] D. Roy and A. Pentland. Learning words from natural audio-visual input. In *International Conference of Spoken Language Processing*, December 1998.

[12] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[13] K. Sung and T. Poggio. Example-based learning for view-based human face detection. Technical Report A.I. Memo 1521, MIT A.I. Lab, 1994.

[14] P. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 1995.

[15] L. Xu, A. Krzyzak, and C. Suen. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3):418–435, May/June 1992.