

Context-Driven Model Switching for Visual Tracking

Hannes Kruppa, Martin Spengler, Bernt Schiele

Perceptual Computing and Computer Vision Group
ETH Zurich, Switzerland
email: {kruppa,spengler,schiele}@inf.ethz.ch

Abstract. A major challenge for real-world object tracking is the dynamic nature of the environmental conditions with respect to illumination, motion, visibility, etc. For such an environment which may experience drastic changes at any time, integration of multiple and complementary cues promises to increase robustness of visual tracking. Nevertheless, one has to expect that false positive tracking will occur. In order to be able to recover from such tracking failure this paper introduces a novel method for automatically choosing the object model which best fits the current context based on information-theoretic concepts. In order to validate the effectiveness of the proposed model switching, it is integrated into a multi-cue face tracking system and experimentally evaluated.

1 Introduction

Many computer vision algorithms have proven to work well in controlled environments or specific contexts. Most often however, real-world environmental conditions are anything but controlled or known a priori. In that sense computer vision algorithms such as tracking should be able to deal with dynamic and even dramatic changes of the environment. In the context of robotics outdoor environments pose a serious challenge to most existing computer vision algorithms. But even in the case of indoor environments unpredictable and drastic changes may occur when the robot moves between different rooms, when the illumination varies depending on weather conditions or when lights are being turned off or on. Other dynamic changes with respect to pose, motion, or occlusion are similarly demanding. In order to deal with dynamic changing environments and with false positive tracking the paper proposes a technique for selecting the model most appropriate for the current context. The ultimate goal of the proposed approach is to overcome the limitations of a set of individual models by selecting the best-suited model on-the-fly, using information-theoretic concepts. To evaluate the effectiveness of the method, the approach is integrated into a tracking system. As an example application we consider tracking of a walking human. In that scenario abrupt direction changes of the walking human and sudden illumination changes may cause tracking failure with which the system has to deal.

The remainder of the paper is organized as follows: section 2 reviews related work, section 3 introduces the concept of mutual information and how it is used in the proposed model switching scheme. As an application a multi-cue human face tracking system is introduced in section 4. Experimental results show that the proposed model switching scheme increases the robustness of the system with respect to false positive tracking in dynamically changing environments. Conclusions and future work are discussed in section 5.

2 Related Work

There seems to be a general consensus that it is important to integrate information coming from different sensors and models in order to increase robustness of today's computer vision algorithms. Whereas in robotics sensor fusion is a common research topic relatively little research has been done in computer vision. Toyama and Hager [TH96] for example propose a layered hierarchy of vision based tracking algorithms aiming to enable robust and adaptive tracking in real-time. When the conditions are good, an accurate and precise tracking algorithm is chosen and when conditions deteriorate more robust but less accurate algorithms are chosen. Eklundh et al. [UNME95, MEN96] argue for a system oriented approach to fixation and tracking in which multiple cues are used concurrently. The important visual cues are motion and disparity which enable to track and fixate single targets using a stereo camera head. Isard and Blake propose an algorithm called CONDENSATION [BI98] which is well suited to track multiple target hypotheses simultaneously. The original algorithm has been extended [IB98a] by a second cue (color in their case) which allows to recover from tracking failures. Yet another variant of this algorithm allows switching between two or three different *motion* models, each of which is defined manually [IB98b]. Other approaches use HMMs to model complex motion patterns, for example to recognize sign language [SWP98].

Even though some of these algorithms integrate information from different visual cues they all rely on the implicit assumption that the context and the environmental conditions are either constant or change relatively continuously. However, when that assumption is violated one has to expect that false positive tracking will occur. To robustly deal with a changing environment the approach proposed in this paper switches between *detection* models based on the current context. Since previous model switching work focused solely on motion models a novel criterion suitable for detection models had to be derived: It is to the best knowledge of the authors the first approach to automatically switch between detection models according to context. *Context-driven model switching* evaluates candidate models based on their soundness within the *current context*. In particular the approach finds the model which best fits certain a priori expectations. Such expectations may be given as constraints on the form of the output distribution which is detailed in the following section.

3 A Mutual Information Criterion for Switching between Detection Models

For visual tracking in the real world, one would like to choose the model that best-matches the current context. To this end, the proposed approach maintains a set of candidate detection models each of which models the object of interest within a distinct environmental condition. That is, these models have been learned in a previous stage using, for example, the maximum likelihood estimator and their parameters remain fixed in our selection scheme. Applying each candidate model to the current scene yields a set of distinct *output distributions* $p(x, y|\theta_i, I)$ where θ_i denotes the n -dimensional parameter vector of the i -th model and I is the current input frame. For every subregion of I , $p(x, y|\theta_i, I)$ is the probability that the target object is contained in the subregion at (x, y) .

The key idea is that if a model performs well within the current context, then $p(x, y|\theta_i, I)$ will have certain distinguishable properties. These exact properties can be tested by comparison to another distribution $p(x, y|\theta^*)$ associated with a *reference model* θ^* . The actual parameter values of θ^* are, of course, unknown. Typically, however, we can learn and constrain the form of appropriate output distributions. For example in the context of frontal face detection we expect a distribution of elliptical shape. This leads to the definition of a parameter space for permissible output distributions. By rating each candidate model θ_i by comparing $p(x, y|\theta_i, I)$ to permissible $p(x, y|\theta^*)$ we can find the model which is in maximal agreement with the expected form of the output distribution in the current context. If on the other hand the candidate model θ_i does not match the current environmental conditions, then we can expect that the form of $p(x, y|\theta_i, I)$ will deviate significantly from $p(x, y|\theta^*)$.

In this paper, we propose to use mutual information as the scoring function for model switching. Mutual information has been used previously in computer vision, for example in medical image registration [Vio95], for selection of most discriminant viewpoints of objects, as well as for the combination of object models [KS00]. As detailed below, mutual information can be used to measure the *mutual agreement* between two distributions. The mutual information of two random variables U and V with a joint probability mass function $p(u, v)$ and marginal probability mass functions $p(u)$ and $p(v)$ is defined as [CT91]:

$$I(U; V) = \sum_{u_i, v_j} p(u_i, v_j) \log \frac{p(u_i, v_j)}{p(u_i)p(v_j)} \quad (1)$$

To undermine the relevance of mutual information in this context, we briefly refer to the well-known Kullback-Leibler divergence. The KL-divergence between a probability mass function $p(u, v)$ and a distinct probability mass function $q(u, v)$ is defined as:

$$D(p(u, v)||q(u, v)) = \sum_{u_i, v_j} p(u_i, v_j) \log \frac{p(u_i, v_j)}{q(u_i, v_j)} \quad (2)$$

The Kullback-Leibler divergence (also called relative entropy or information divergence) is often used as a distance measure between two distributions. By defining $q(u, v) = p(u)p(v)$ the mutual information can be written as the KL-divergence between $p(u, v)$ and $p(u)p(v)$:

$$I(U; V) = D(p(u, v)||p(u)p(v)) \quad (3)$$

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

Let M_i and M^* be a pair of discrete random variables with $M_i, M^* \in \{0, 1\}$. We define

$$\begin{aligned} p(M_i = 1|\theta_i, I) &= \int_{x, y} p(x, y|\theta_i, I) dx dy & p(M^* = 1|\theta^*) &= \int_{x, y} p(x, y|\theta^*) dx dy \\ p(M_i = 0|\theta_i, I) &= \int_{x, y} (1 - p(x, y|\theta_i, I)) dx dy & p(M^* = 0|\theta^*) &= \int_{x, y} (1 - p(x, y|\theta^*)) dx dy \end{aligned}$$

	$M_i = 1$	$M_i = 0$
$M^* = 1$	$\frac{1}{\eta} \int_{x,y} p(x, y \theta_i, I) * p(x, y \theta^*) dx dy$	$\frac{1}{\eta} \int_{x,y} (1 - p(x, y \theta_i, I)) * p(x, y \theta^*) dx dy$
$M^* = 0$	$\frac{1}{\eta} \int_{x,y} p(x, y \theta_i, I) * (1 - p(x, y \theta^*)) dx dy$	$\frac{1}{\eta} \int_{x,y} (1 - p(x, y \theta_i, I)) * (1 - p(x, y \theta^*)) dx dy$

Table 1.: Definitions of $p(M_i, M^*|\theta_i, I, \theta^*)$ used for computing mutual information. $\frac{1}{\eta}$ is a normalizing constant

such that M_i and M^* indicate the presence or absence of the modeled object according to $p(x, y|\theta_i, I)$ and $p(x, y|\theta^*)$, respectively. Further, table 1 lists definitions of the joint distribution $p(M_i, M^*)$. For the moment we assume that a reference model θ^* is given with a specific number of parameters, each of which can be associated with a range of permissible values. A practical example how to derive these constraints will be given in the next section. Now, for each candidate model θ_i the algorithm searches the configuration space of θ^* , maximizing their mutual information $I(M_i; M^*|\theta_i, I, \theta^*)$

$$\max_{\theta_i, I, \theta^*} I(M_i; M^*|\theta_i, I, \theta^*) = \sum_{M_i, M^* \in \{0,1\}} p(M_i, M^*|\theta_i, I, \theta^*) \log \frac{p(M_i, M^*|\theta_i, I, \theta^*)}{p(M_i|\theta_i, I)p(M^*|\theta^*)} \quad (4)$$

The model θ_0 of $\{\theta_i\}$ which reaches the highest mutual information score is then chosen.

3.1 Example: Switching between Skin Color Models

We applied the above approach to automatically choose between skin color models in different lighting situations. In a first step, we trained a set of five distinct Gaussian models θ_i using the Maximum Likelihood Estimator for modeling Caucasian skin color in HSI color space. For the training, five separate data sets gathered from distinct indoor and outdoor situations were used (bright sun, cloudy, lab, candlelight, green filtered light) to represent different contexts. In the proposed scheme we now need to derive the parameter number as well as parameter ranges which constrain the reference model θ^* . Essentially, these constraints need to capture properties of $p(x, y|\theta_i, I)$ which allow to evaluate the appropriateness of a candidate model θ_i in the current context I . While it might be possible to learn such properties from data, it is in this case straight-forward to come up with a direct definition: As skin appears as *contiguous patches* of approximately elliptical shape (faces, hands, limbs), the reference distribution $p(x, y|\theta^*)$ can be defined as a unimodal distribution with a local elliptical plateau of high probability. That means we have $\theta^* = (x_c, y_c, w, h)$, where (x_c, y_c) denotes the plateau’s center within the distribution, and (w, h) are the plateau’s dimensions. The actual probabilities come from the *logistic function*

$$\text{logistic}(a) = \frac{1}{1 + \exp^{-a}} \quad (5)$$

We choose values of a such that probabilities within the plateau become relatively high and adapt the parameter towards the plateau’s boundary so that probabilities decrease smoothly and eventually reach zero. While our current implementation defines only upright elliptical regions, the extension to tilted regions would be straight-forward, but incurs the cost of increased model complexity of θ^* . For the computation of expression (4) the algorithm traverses (x, y) -space starting at the global maximum (x_0, y_0) of $p(x, y|\theta_i, I)$ and then visiting local maxima in descending order. At every location, the algorithm performs a local search in (w, h) -space, evaluates (4) and stores the local maximum mutual information value. For further speed up, one can restrict the search of (x, y) such that only $k\%$ of all locations are examined. Likewise, it may be possible to impose additional value constraints on (w, h) , depending on the application in mind.

Table 2 shows a set of test images taken under different lighting situations. Below each image a histogram of mutual information for the described skin color models is shown (for clarity, logarithms of mutual information are given in the table). The maxima of mutual information (highlighted in the histograms) suggest that mutual information may be used to choose the appropriate models automatically.

4 Application: Robust Tracking of Human Heads

In order to test the context-driven model switching scheme proposed in the previous sections, this section describes its application to people tracking. In particular, we extended an existing multi-cue head tracking system [SS] with context-driven switching between skin color models.

One important problem of visual tracking in general and head tracking in particular are sudden changes in the perceived environment. Unexpected and unpredictable situations are likely to disturb even highly sophisticated feature detectors used in state-of-the-art visual tracking systems, decreasing the system’s reliability. As already mentioned

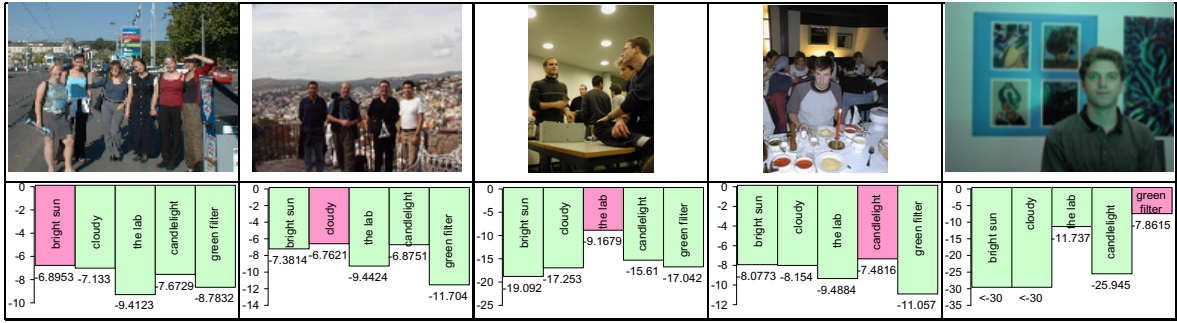


Table 2.: The table shows the logarithm of the mutual information of the pictures at the top with five distinct color models which model Caucasian skin color in different contexts (bright sun, cloudy, lab, candlelight, green filtered light). The mutual information values are visualized as histograms and the maxima are highlighted. These mutual information maxima suggest that an appropriate color model can be chosen based on mutual information

in the introduction, one strategy to cope with such disturbances is to integrate multiple visual cues or observations respectively. The underlying idea is that parallel use of multiple discriminate information channels increases the system’s overall reliability by covering a wider feature range. That is, the probability of observing a feature not influenced by a particular change in the environment increases. The head tracking system we use for the experimental verification of context-driven model switching implements a concept of self-organized sensor fusion originally proposed by Triesch et. al [Tvdm00] called *Democratic Integration*: The head position estimates of several auto-adaptive visual cues like motion detectors, color detectors, or template matchers are merged in order to obtain a robust position estimation based on mutual agreement. To increase robustness sensor integration scheme adapts dynamically and depending on he context.

Experiments with the head tracker proved the idea of self-organized cue integration as a powerful concept to achieve robust and reliable tracking [SS]. Although capable of coping with several severe context disturbances like abruptly changing ambient illumination, our experiments raised two significant shortcomings of *Democratic Integration*: First, the integration mechanism is not able to *bootstrap* failing cues. That is, failures due to sudden environmental changes can only be recovered by steady but usually slow adaptation of the applied model and not by resetting the cue to a new, more appropriate model. Therefore a failing cue will be suppressed for a longer period, decreasing the system’s overall reliability. Second, *false positive tracking* causes “dead-locks” of the tracking system upon wrong hypotheses. Whenever a hypothesis is accepted (correct or false), the system begins to adapt itself to this hypothesis, reinforcing it even more. In order to escape from dead-locks, the tracking system depends on an external reset mechanism.

As will be shown in subsection 4.2 below, context-driven model switching implements a boot-strapping mechanism for visual cues. By on-demand resetting of the applied model, the cue is able to react immediately on an altered context as well as to recover from false positive tracking.

4.1 Democratic Integration

In order to detect and to track a person in a video sequence, *Democratic Integration* fuses the position estimations of several visual cues. Every visual cue is dedicated to a particular observation, e.g. intensity change, skin color detection, shape template matching, contrast, etc. Multi-cue integration is formulated as follows:

$$p_C(\mathbf{x}|t) = \sum_i \omega_i \cdot p_i(\mathbf{x}|t) \quad (6)$$

where $p_C(\mathbf{x}|t)$ denotes the combined head position estimation the cues agreed upon at time step t . A particular visual cue i contributes the probability distribution $p_i(\mathbf{x}|t)$ which is weighted with its reliability $\omega_i \in [0, 1]$. These weights are dynamically adapted in order to reflect a cue’s current performance or reliability respectively.

The estimated head position $\hat{\mathbf{x}}(t)$ is defined as the maximum response of the combined probability distribution $p_C(\mathbf{x}|t)$:

$$\hat{\mathbf{x}}(t) = \arg \max_{\mathbf{x}} \{p_C(\mathbf{x}|t)\} \quad (7)$$

Once this target position is estimated, the visual cues can be evaluated and their weights adapted accordingly. The correlation of a cue’s estimation $p_i(\mathbf{x}|t)$ and the combined estimation $p_C(\mathbf{x}|t)$ defines a quality measurement $\tilde{q}_i(t)$ for the cue’s current performance. The relation of weight $\omega_i(t)$ and quality $\tilde{q}_i(t)$ is then given by the following dynamics:

$$\tau_i \cdot \dot{\omega}_i(t) = \frac{\tilde{q}_i(t)}{\sum_j \tilde{q}_j(t)} - \omega_i(t) \quad (8)$$

where τ_i is a time constant determining how fast the weight ω_i is adapted. Quality $\tilde{q}_i(t)$ is normalized in order to guarantee $\sum_i \omega_i(t) = 1$ at any time. Besides this self-organization property of *Democratic Integration*, the visual cues themselves may be auto-adaptive. That is, a cue's object model may be adapted in order to cope with fluctuations in the environmental context. Self-adaptation of visual cues further increases the robustness of self-organized multi-cue integration.

4.2 Experimental Results

In order to emphasize context-driven model switching's ability to bootstrap the skin color detector and thus preventing false positive tracking, two exemplary tracking sequences will be discussed below. In the following we describe two different modes of the system: in the first mode *Democratic Integration* uses a standard skin color detector and in the second mode *Democratic Integration* uses skin color detection with the model switching scheme as proposed in the previous sections.

Besides the skin color detection cue, both systems consists of three additional visual cues observing the following features: intensity change between two subsequent images, similarity to a given head template, and contrast. A fifth visual cue predicts the targets motion with respect to its past trajectory. Initializing the reliabilities of the intensity change cue and the skin color detection cue to 0.5 expresses the a priori knowledge that both systems are supposed to track skin colored regions which are in motion. The initial reliabilities of the remaining three cues are set to 0.

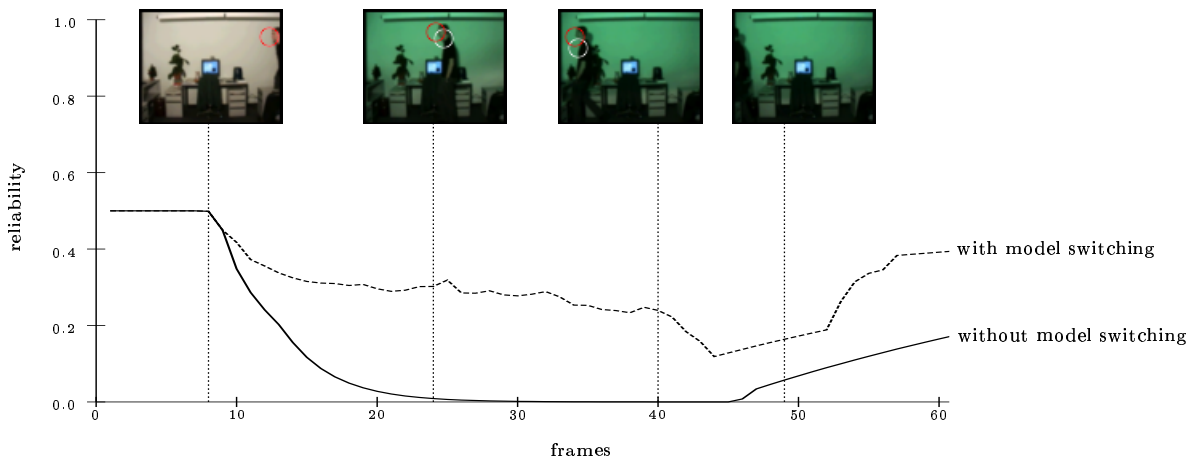


Fig. 1.: **Color-change sequence.** From frame 24 the ambient light is abruptly changed by a green filter. For this situation, context-driven model switching automatically chooses a suitable skin color model. Hence the reliability of the skin color detector (y-axis) is stabilized (note that reliabilities of all cues involved always sum up to one) and the target hypothesis is accurate (red circle). Without model switching tracking fails (white circle)

Sequence 1: Color change In this sequence a subject crosses the scene from right to left. While the subject is in the scene the ambient illumination abruptly changes its color from white to green, challenging the system's ability to cope with large alternations of the environmental context.

When the subject enters the scene in frame 8, both systems begin to converge toward an equilibrium (figure 1). That is, the reliability of intensity and color is decreased whereas the reliabilities of the remaining cues increase. Due to accidental false positive tracking, the conventional tracker begins to converge toward a wrong hypothesis, leading to complete failure of the color cue. This early collapse of one of the two crucial visual cues prevents the whole system from establishing a stable tracking hypothesis. These instabilities hinder the color cue to extract a reasonable skin color model resulting in a suppression of the color cue until the subject has left the scene and tracking is lost (figure 1).

In contrast, the second system using a color cue with context-driven model switching resets the applied skin color model repeatedly, preventing the whole system to lock on a false positive (figure 1). As a consequence, the system

overcomes accidental false positive tracking and is capable to track the subject's head correctly. Stabilized by its boot-strapping mechanism, the color cue not only manages to survive the radical alternation of the environmental context without any particular loss of reliability but is even capable to compensate the intensity cue's temporary failure (figure 1, peak at frame 25).

The rise of the color cue's reliability after the subject has left the scene (figure 1, frame 53) is caused by false positive tracking triggered by the re-adaptation of the skin color model to its default value. Long-time re-adaptation to default values in absence of any potential target is vital for a system like ours in order to reset the system to a defined initial state.

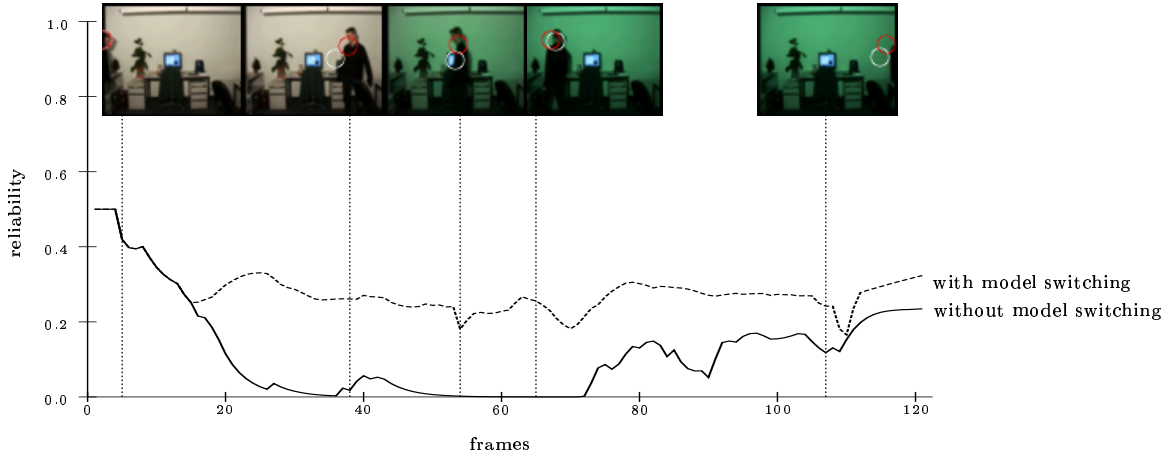


Fig. 2.: **Color change and turning sequence.** Without model switching, the systems adapts to the background and hence tracking fails (reliability plot, frame 18). In contrast, automatic switching of models makes the tracking robust even with the person turning (red circle frame 39) and also deals successfully with the abrupt change in lighting in frame 53. Here, the original tracker fails again.

Sequence 2: Color change and turns In sequence two, the subject enters the scene at the left side, walks over to the right side, turns (face to camera), walks back to the left, turns again (face away), and walks to the right where he leaves the scene. Between the first and second turn, the ambient illumination is changed from white to green.

As illustrated in figure 2, both systems initially converge towards an equilibrium when the target is tracked. At frame 18 false positive tracking occurs. The color cue of the original system fails leading to similar effects as described for the first sequence: Instabilities preventing the system's recovery. The improved system with the skin color detector using model switching is able to overcome this false positive tracking. Compared with the original color cue, the extended cue is more reliable and stabilizes the head tracking system throughout the whole sequence.

The effect of boot-strapping is again illustrated at frame 53 in figure 2: The color cue's reliability drops due to the abrupt change of the ambient illumination. The skin color detector is boot-strapped immediately after its failure and regains its former reliability within two frames. After altering the actual color model to the selected one, self-adaptation starts and tunes this coarse color model according to the perceived context.

The analysis of the two different head tracking systems shows that context-driven model switching improves the performance of the skin color detection cue as well as the stability of the whole tracking system. As claimed at the beginning, the switching scheme is able to bootstrap a visual cue, accelerating the cue's ability to adapt to a new environment. Furthermore, it helps the system to recover from false positive tracking resulting in an improved reliability and robustness of the whole tracking system.

5 Conclusion and Future Work

This paper introduced *context-driven model switching* where the current context is used to choose from a set of previously learned object models. More specifically, the validity of a model in the current context is tested by comparing the resulting output distribution to an a priori expectation. We derive a mutual information criterion for this comparison and show that the object model which maximizes mutual information also best-matches the current context. We also derive a reference distribution for switching between different skin color models. The approach itself however is not limited to color models but is applicable to any type of visual cue. As an example application, we

apply the method to overcome false positive tracking in a multi-cue face tracker. The experimental results indicate that the system's performance can be leveraged by using context-driven model switching.

Currently we are running additional experiments to investigate model switching where models are based on other kinds of visual cues. For the case of template based models, for example, preliminary results indicate that a useful switching criterion can be established using a slightly modified reference distribution. Also, we compare the relative detection performance to other skin color models [SAG99] as well as different face models [SK00, RBK98].

References

- [BI98] Andrew Blake and Michael Isard. *Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer, 1998.
- [CT91] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, Inc, 1991.
- [IB98a] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *ECCV'98 Fifth European Conference on Computer Vision, Volume I*, pages 893–908, 1998.
- [IB98b] M. Isard and A. Blake. A mixed-state condensation tracker with automatic model-switching. In *ICCV'98 Sixth International Conference on Computer Vision*, pages 107–112, 1998.
- [KS00] H. Kruppa and B. Schiele. Using mutual information to combine object models. In *8th International Symposium on Intelligent Robotic Systems'00*, pages 311–316, July 2000.
- [MEN96] A. Maki, J.-O. Eklundh, and P. Nordlund. A computational model of depth-based attention. In *International Conference on Pattern Recognition*, 1996.
- [RBK98] 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.
- [SAG99] Moritz Stoerring, Hans J. Andersen, and Erik Granum. Skin colour detection under changing lighting conditions. In *7th International Symposium on Intelligent Robotic Systems'99*, pages 187–195, July 1999.
- [SK00] 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.
- [SS] Martin Spengler and Bernt Schiele. Towards robust multi-cue integration for visual tracking. submitted to ICVS 2001.
- [SWP98] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1372–1375, 1998.
- [TH96] K. Toyama and G. Hager. Incremental focus of attention. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1996.
- [TvdM00] J. Triesch and C. von der Malsburg. Self-organized integration of adaptive visual cues for face tracking. In *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 102–107, March 2000.
- [UNME95] T. Uhlin, P. Nordlund, A. Maki, and J.-O. Eklundh. Towards an active visual observer. In *ICCV'95 Fifth International Conference on Computer Vision*, pages 679–686, June 1995.
- [Vio95] P. Viola. *Alignment by Maximization of Mutual Information*. PhD thesis, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, 1995.