IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications
21-23 September 2009, Rende (Cosenza), Italy

Coarse-to-fine Particle Filters for Multi-Object Human Computer Interaction

Matthias Rätsch[1, 2], Clemens Blumer[1], Gerd Teschke[2], and Thomas Vetter[1]
[1] University of Basel, Bernoullistrasse 16, CH-4057 Basel, Switzerland,
{matthias.raetsch, clemens.blumer, thomas.vetter}@unibas.ch, http://gravis.cs.unibas.ch
[2] University of Applied Sciences Neubrandenburg, Brodaer Str. 2, D-17033
Neubrandenburg, Germany, {raetsch, teschke}@hs-nb.de, http://www.hs-nb.de

***Abstract* – Efficient motion tracking of faces is an important aspect for Human Computer Interaction (HCI). In this paper we combine the Condensation and the Wavelet Approximated Reduced Vector Machine (W-RVM) approach. Both are joined by the core idea to spend only as much as necessary effort for easy to discriminate regions (Condensation) or vectors (W-RVM) of the feature space, but most for regions with high statistical likelihood to contain objects of interest. We adapt the W-RVM classifier for tracking by providing a probabilistic output. In this paper we utilize Condensation for template based tracking of the three-dimensional camera scene. Moreover, we introduce a robust multi-object tracking by extensions to the Condensation approach. The novel coarse-to-fine Condensation yields a more than 10 times faster tracking than state-of-art detection methods. We demonstrate more natural HCI applications by high resolution face tracking within a large camera scene with an active dual camera system.**

***Keywords* – Human Computer Interaction, Multi-Object Face Tracking, Condensation, Coarse-to-fine Particle Filters, Wavelet Approximated Reduced Vector Machine, Active Dual Camera System**

I. INTRODUCTION

Faces always have a high attraction to humans, they are able to predict immediately the position, movements, or expressions of faces. Eye-contact is an important aspect for non-verbal interaction in the field of perception psychology. In the future, Human Computer Interaction (HCI) should be as natural as a conversation between humans. An embodied conversational agent or humanoid robot must be able to localize its 'conversational partner' before it can get in contact. A machine which can detect objects is an important aspect of computer science. Especially that a machine can localize a humans face and interact in some manner with the person is a fascinating issue.

Image-based detection tasks are time consuming. For instance, detecting a specific object in an image, such as a face, is computationally expensive, as all pixels of the image are potential object centers. Hence, all pixels must be classified, for all possible object sizes. The fastest state-of-the-art classifiers, for example the AdaBoost based classifier of Viola and Jones [1] or the Wavelet Reduced Vector Machine introduced by Rätsch et al. [2], are applied to detection algorithms near real-time. Detection uses a sliding observation window strategy. The brute-force search cuts out patches and classifies them for each pixel location of the entered image. To detect objects of different size (i.e. objects at different distances to the camera) an image pyramid is used by down-sampling the image several times till the object has the size of the observation window. However, for video streams with high-resolution cameras, covering a large range of distances between the camera and the object, or/and if we want to detect different object classes at the same time (e.g. facial features like eyes, nose tip, and mouth corners) the sliding observation window strategy quickly becomes intractable.

It is obvious that the object's position and size vary only slightly from one video frame to the next. Therefore, it is possible to use information from the last time steps to speed up the search in the next frame. The process of seeking and following objects is called tracking. A method that is capable of using information of the previous iterations is the Condensation algorithm and was proposed by Isard and Blake [3]. Condensation is able to track objects in a highly cluttered background. The tracking method is a good alternative to the Kalman Filter [4], because Condensation can estimate the unknown a-posteriori probability function and does not need the assumption of a Gaussian distribution. Therefore, the estimated density function is multi-modal (i.e., it can have several maxima). The system and measurement dynamics can be nonlinear and they are suited for parallelization. The original Condensation approach by Isard and Blake is introduced to track contours of objects. We adapted the approach for tracking objects using template based classifiers. In this paper we propose to combine Condensation tracking with our efficient Wavelet Reduced Vector Machine (W-RVM) [2, 5, 6]. The W-RVM uses a Double Cascade for early rejections of easy to discriminate image locations. The classifier gains a more than 500 fold speed-up compared to an original Support Vector Machine [7]. The classifier trains much faster as the Viola and Jones classifier [1] by same detection accuracy and run-time performance and detects about 25 times faster than the Rowley-Baluja-

This work was supported by a grant from the German Research Foundation (DFG TE 354/4-1).

Kanade detector [8] and about 1e3 times faster than the Schneiderman-Kanade [9] detector. The novel Cascaded Condensation Tracking (CCT) unifies the core ideas of the Condensation and W-RVM approach to spend less computational effort for easy to discriminate feature space locations. Instead measuring each pixel of the frame Condensation contracts particles at areas with higher interest. Additionally, the W-RVM spends at each of these feature space locations of the particles only as much as necessary effort by adapting the core-to-fine Double Cascade to the tracking approach and refining the measurement step of the Condensation approach. The drawback of multi-modal Condensation is that it cannot track stably multiple objects over a longer time period. Kang et al. [10] changed the Condensation algorithm to be usable with multiple objects of the same class, e.g. faces. The main idea is to build multiple trackers which are in concurrence and hold only their main area. By Kang's approach for every object a tracker instance (with an own set of particles) is needed. So the number of trackers depends on the number of objects detected. In difference, our approach will take advantage of the multi-modal density function of Condensation. We will use one tracker with a single set of multi-modal particles which handles the different objects of the same class. As novelty we also introduce a minimal density constraint for robust multi-object tracking. The next limitation of tracking approaches is that they are limited to track only the in-plane translations of objects (x- and y-coordinates) and cannot be used for other feature vectors or higher dimensions, e.g. the object distance to the camera as a third tracking dimension. Bretzner et al. [11] propose a specialized multi-scale tracking like for features different in size or Yang et al. [12] and Huang et al. [13] use specific deformable templates. In contrast, we want to introduce a novel abstract multidimensional feature vector tracking, able to distribute the density function of the particles over higher dimensional abstract feature vectors. For example our approach will be applied for the three-dimensional Condensation tracking of the x-, y-, and z-coordinates of objects, where the z-dimension is the distance of the object to the camera [14]. Our approach will be open for tracking abstract feature vectors and with more than three dimensions, e.g. the orientation of the objects or even abstract object or model parameters. If faces and other facial features (e.g. eyes) can be tracked stably, in real-time, and over larger distances Human Computer Interactions become much more natural because the interaction area is larger and more convenient. Current systems mostly track faces only over low distances, e.g. sitting in front of a camera. Moreover, for most facial applications only high resolution images are suitable. For example, to apply the 3D Morphable Face Model (3DMM, [15]) for face or facial emotion recognition, we want to use a dual camera system with a static and a Pan-Tilt-Zoom (PTZ or active) camera which can be rotated and optical zoomed. Prince et al. [16] propose a dual camera system to deliver high resolution images. In the static image the detection is based on background subtraction and the skin/background-color of the body. They direct the active camera on a face and apply a face recognition system on the image section. In difference to them, we will detect and track faces alternatively on the static or active camera for most robust tracking [17]. By Yang et al. [12] an approach with an active camera was realized. They do a detection based on color combined with an online learning. To detect new faces beside the online learning model a face detector is used. It is not clear stated if the detector is only based on color information. Our approach will use a powerful classifier based on the double cascaded W-RVM, using a Support Vector Machine as final validation stage, known for best generalization performance [7]. It is not detailed if Yang et al. use zoom facilities in case an object is detected. So their system seems not able to provide high resolution images of faces at larger distances. Summarizing our approach, we will introduce inventive extensions for the Condensation tracking by Isard and Blake and unify the approach with our efficient double cascaded coarse-to-fine W-RVM classifier. The core ideas of the W-RVM classifier and its extension for probabilistic outputs are summarized in Section II. The novel Cascaded Condensation Tracking (CCT, Section III) joins the core idea of both approaches to spend less computational effort for easy to discriminate image regions (Condensation) and vectors (W-RVM) of the feature space, but most for locations with high statistical likelihood to contain the object of interest. Moreover, we adapt Condensation from tracking curves to general density functions over abstract multidimensional feature vectors suitable for template based classifiers. We apply the multidimensional CCT by distributing the particles over the two in-plane translation coordinates and additional over the distance of the object to the camera as third dimension in the 3D feature vector. Our 3D CCT is also able to track stably multiple objects with one single set of particles by an adaptive multi-modal probability distribution, a weighted drift function, and a minimal density constraint (Section III-B). The proposed approach can be used for any kind of objects. For our experiments we use human faces or facial features like eyes. We apply the new CCT approach to a PTZ-camera and an active dual camera system and will demonstrate HCI applications 10 times more efficient as state-of-the-art detection methods (Section IV).

II. Probabilistic Wavelet Approximated Reduced Vector Machine

Face detection or detection in general is the process to search for a specific object-class (e.g. faces) and locate the object in images. The goal is to classify a given image point with a given patch size as object or non-object. This means object detection is a binary pattern-classification problem. Face detection is complex as faces

differ in size, rotation, pose and illumination. Furthermore, glasses often occlude parts of the characteristic eyes and specular highlights occur. By classification of every image point as potential object center and for all possible object sizes for a standard VGA frame over 3e5 classifications are needed. With for example 1ms per classification more than five minutes would be needed to process a single frame. Therefore, reductions in classification and number of sample points are required. The reduction of used sample points is gained by tracking and the reduction per sample by the coarse-to-fine W-RVM measurement.

We will now roughly introduce the core ideas of the Wavelet Approximated Reduced Vector Machine (W-RVM) and how to obtain a probabilistic measurement output. The W-RVM classifier is a two stage approximation of a Support Vector Machine (SVM). Suppose that we have a labeled training set consisting of a series of e.g. 20×20 image patches $\mathbf{x}_i \in \mathcal{X}$ (arranged in a 400 dimensional vector) along with their class labels $y_i \in \{\pm 1\}$. Support Vector classifier implicitly map the data $\mathbf{x}_i$ into a dot product space F via a (usually nonlinear) map $\Phi : \mathcal{X} \rightarrow F, \quad \mathbf{x} \mapsto \Phi(\mathbf{x})$. Although F can be high dimensional, it is usually not necessary to explicitly work in that space [7]. By Mercer's theorem, it is shown that it exists a class of kernels $k(\mathbf{x}, \mathbf{x}')$ to compute the dot products in associated feature spaces, i.e. $k(\mathbf{x}, \mathbf{x}') = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$. The training of an SVM provides a classifier with the largest margin [7], i.e. with the best generalization performances for given training data and a given kernel.

The following core ideas of the W-RVM provide an optimal approximation of the decision hyper-plane for an efficient and accurate classifier (For more details about the used classifier, we refer the reader to [2, 5, 6].):

1) **Support Vector Machine:** Use of an SVM [7] classifier that is known to have optimal generalization capabilities.
   a) SVM: $\Psi_{\text{SVM}} = \sum_{i=1}^{N_x} \alpha_i \, \Phi(\mathbf{x}_i)$, $\mathbf{x}_i$ are the Supports Vectors (SSV's)
   b) Decision function: $y(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{N_x} \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b\right)$ with the kernel function $k(\cdot, \cdot)$, e.g. Gaussian kernel $k(\mathbf{x}, \mathbf{x}_i) = \exp(-||\mathbf{x} - \mathbf{x}_i||^2/(2\sigma^2))$.
2) **Reduced Support Vector Machine:** The SVM is reduced by a set of Reduced Set Vectors (RSV's, $\mathbf{z}_i$) [18]. Fig. 1 shows on a 2D toy example that with only 9 RSV's instead of 31 SSV's ($N_z \ll N_x$) the same decision accuracy can be obtained.
   a) RVM: $\Psi_{\text{RVM}} = \sum_{i=1}^{N_z} \beta_i \, \Phi(\mathbf{z}_i)$.
   b) Decision function: $y(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{N_z} \beta_i \, k(\mathbf{x}, \mathbf{z}_i) + b_i\right)$.
3) **Double Cascade:** The RVM is reduced in a second step by approximating each RSV by several levels of Wavelet Approximated Reduced Set Vectors (W-RSV's). For non-symmetric data (i.e. only few positives to many negatives) we achieve an early rejection of easy to discriminate vectors using a Double Cascade over coarse-to-fine W-RSV's:
   a) **Cascade over the number of used W-RSV's**: Using only the first reduced vectors yields high error rates (Fig. 1), but data points (with a large negative distance to the classification boundary) can be early rejected as negative points, without further evaluation cost.
   b) **Cascade over the resolution levels of each W-RSV**: Already using the first approximations stages of the 2nd cascade (e.g., Fig. 2, *left to right*), first image locations, like homogenous background, can be rejected. Only for more difficult image locations the full complexity of the W-RSV's must be used.

   The Double Cascade constitutes one of the major advantages of the W-RVM approach. The trade-off between accuracy and speed is very continuous.
4) **Integral Images:** As the W-RSV's are approximated using a Haar wavelet transform, the Integral Image method is used for their evaluation [5].
5) **Wavelet Frame:** We use an over-complete wavelet system to find the best representation of the W-RSV's. The learning stage of the W-RVM is fast, automatic, and does not require the manual selection of ad-hoc parameters. For example, the training time is about two hours [6], instead in the order of weeks like the Viola and Jones classifier [1]. The Over-Complete Wavelet Transform is applied at the W-RVM training. That is opposite to several other approaches using a wavelet input space transformation as a pre-processing at detection time.

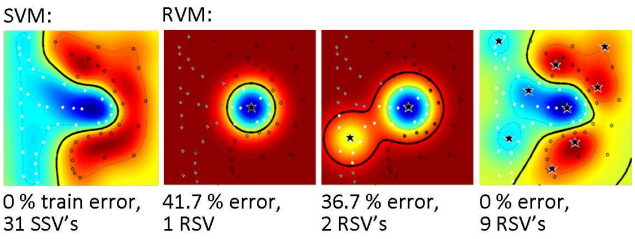

Fig. 1. Cascaded application of RSV's (*stars*) to a 2D classification problem (*black and white dots*), showing (*left to right*) the original SVM and the result of using 1, 2, and 9 Reduced Set Vectors.

The W-RVM classifiers support binary decision output and a certainty which is related to the distance to the decision hyper-plane. A large distance indicates a higher classification certainty. However, for the Condensation approach probabilistic outputs of the measurement function are needed. We tested for the estimation of the PDF (class-conditional probability) histogram, parzen-window, and k-NN methods, all were not stable enough. Best results we obtained by fitting a sigmoid function for the posterior probability. The sigmoid function fitting is a model-trust

algorithm, based on the Levenberg-Marquardt algorithm [19]. The method extracts probabilities from SVM outputs, which is useful for classification post-processing. The method adds a trainable post-processing step which is trained with regularized binomial maximum likelihood. A two-parameter sigmoid is chosen as the post-processing, since it matches the posterior that is empirically observed.

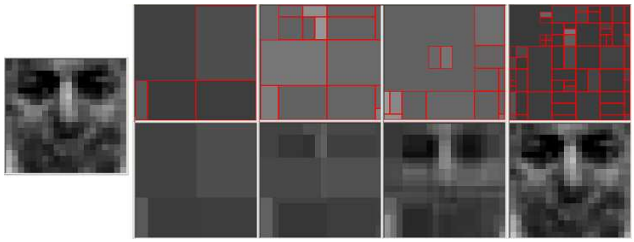

Fig. 2. Example of coarse-to-fine W-RSV's for the first RSV (*left*). W-RSV's at different resolution levels (*bottom row*) and the related wavelet approximated residuals (*above*).

$$p_{ffp}\left(\mathbf{x}_{ffp}|t_{ffp}\right) = \frac{1}{1+\exp\left(A\ t_{ffp}+B\right)} \qquad (1)$$

The sigmoid fitting trains iterative the parameters A and B of the sigmoid function to map the W-RVM output t_{ffp} of the feature point ffp (e.g. faces or eyes) into probabilities $p_{ffp}\left(\mathbf{x}_{ffp}|t_{ffp}\right)$.

III. 3D Cascaded Condensation Tracking for Multiple Objects

A. 3D Cascaded Condensation Tracking

Condensation invented by Isard and Blake [3] stands for 'Conditional Density Propagation' and is one of the most successfully used approaches evaluated for different tracking tasks. This algorithm uses statistics to distribute n samples $\mathbf{s}_i$ with $0 \leq i < n$, by a conditional probabilistic density function over an image and measures only at this certain pixels of the frame if an object of interest is located at these image positions. Instead of all pixels a much lower number n of measurements is needed. This provides a significant speedup. For the measurement function the W-RVM is used.

The initial selection step distributes the set of samples randomly over the full frame. In our approach the samples $\mathbf{s}_i \in \mathbf{X}$ are distributes over all dimensions of the feature vector space $\mathbf{X}$ (e.g., additionally over all scales of the image pyramid (third dimension, $\mathbf{s}_i \in \mathbb{R}^3$)). Then the measurement function evaluates the likelihood $\pi_i = p(\mathbf{z}|\mathbf{x}=\mathbf{s}_i)$ that the object of interest is represent by the image section $\mathbf{z}$ (e.g., a 20×20 grey value patch, $\mathbf{z} \in \mathbb{R}^{400}$) given the sample $\mathbf{s}_i$ is located on the respective feature space location $\mathbf{x}$ for $\mathbf{z}$. The selection step contracts the density function around samples with a high likelihood by selecting more samples at these locations. The following prediction step estimates the drift of the object from the last frames and adds the offset to the samples. Also a Gaussian noise using a diffusion matrix is added to the sample locations at the prediction step. The next loop starts with the measurement function, again. Per time step (frame) one loop over the three steps (selection, prediction and measurement) is performed. For a detailed description, we refer the reader to [3].

B. Tracking of Multiple Objects

An approach able to track multiple instances of the same class of objects (e.g. faces) is substantial for many applications. A drawback of original Condensation is that multi-object tracking is not stable, although it provides a multi-modal density function and probability distribution (function with more than one maximum) as opposite to the Kalman Filter [4]. For the maxima at the density function we use the same clustering approach as in [6], but here by assigning samples to clusters with respect to their weight and Euclidian distance to the cluster centers.

Fig. 3 shows an example for a one dimensional density function with two clusters. Original Condensation contracts more and more samples around the cluster with a higher probability to be an object of interest (*left, cluster 1*). The sample density of the not as probably cluster *cluster 2* is reduced to the background density and is not tracked anymore or a swinging between the objects can result. Only if two clusters would have the exact identical response (what is not the case because of the influence of random values) both would be stably tracked. We propose a novel approach, inspired by Kang [10], but there multiple instances of the tracking method (each with an own set of particles) are used and the advantage of Condensation to provide a multi-modal density function is not exploited.

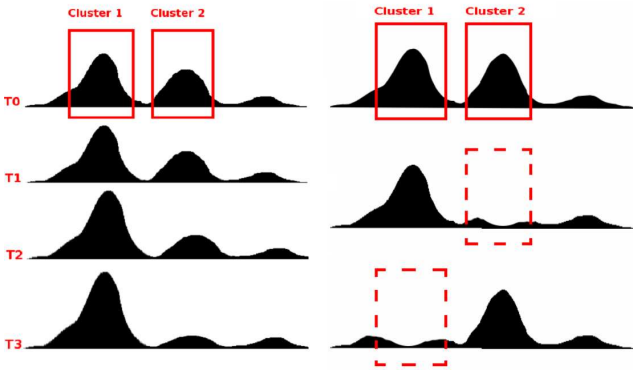

Fig. 3. By original Condensation after some iterations (*left, T0 till T3*) one cluster dominates. Therefore, we adaptively manipulate the probability distribution (*right, top*), so that samples of the other clusters (e.g. samples near cluster 2 (*middle*) or near cluster 1 (*bottom*)) are suppressed.

We use one multi-modal distributed set of samples but adapt the density function individual for every cluster. The original probability distribution is manipulated to suppress samples of other clusters as Fig. 3, *right* shows. No expensive computations are needed. The selection of new samples is modified to use the individual distributions:

- iterate over the adaptive probability distributions with respect to the m clusters $\mathbf{c}_j$ with $0 \leq j < m$:
  - select n/m times a valid sample with each of the m manipulations of the probability distribution.

This manipulation of the selection step of the Condensation method has the effect that n/m samples are used for every cluster and a balance between the different clusters is obtained. For every cluster all certainties are recalculated based on the distances $d_{i,k}$ from the sample to $m-1$ clusters (for the cluster to which the manipulated probability distribution belongs no distance must be calculated). The certainties are multiplied $m-1$ times by a factor $r(d_{i,k}) = 1 - 1/(\exp(d_{i,k}/p)^q)$, where p and q are empirical constants (we obtained good results with $p = 40$ and $q = 6$). The certainties $\pi_{i,j}$ for sample $\mathbf{s}_i$ with respect to cluster $\mathbf{c}_j$ are obtained by:

- calculate the distances to the clusters: $d_{i,k} = \sqrt{(s_{i_x} - c_{k_x})^2 + (s_{i_y} - c_{k_y})^2}$ with $0 \leq k \leq m$ and $k \neq j$.
- multiply π_i (measured likelihood for each $\mathbf{s}_i$) with all $m-1$ factors $r(d_{i,k})$ to compute the likelihood $\pi_{i,j}$

$$\pi_{i,j} = \pi_i \prod_{k=0, k\neq j}^{m-1} r(d_{i,k}) \qquad (2)$$

- finally normalize the new certainties: $\pi_{i,j} = \pi_{i,j} / \sum_{l=0}^{n} \pi_{l,j}$. This process is done for all n samples with respect to all m clusters ($n * m$ new certainties).

The number of objects can be limited (e.g., if only one person is in the image) to c_{max} clusters. To profit from this a-priori knowledge the multi-object certainties are calculated for all found clusters and the best c_{max} clusters are kept. After calculating the weighted certainties the dispensable cluster regions (clusters not in c_{max}) obtain fewer samples at the next iteration and most samples are contracted on the expected clusters.

For multi-object CCT we additionally propose a weighted drift function for the prediction of the next sample positions. This yields a robust tracking, because the objects can move in different directions and with different speed. A linear combination of the 3D offsets of the centers is used for drift of each sample. The weights are evaluated by the Euclidian distance to the cluster centers normalized by the number of clusters. That means the drift of a sample is continually most influenced by the drift of the nearest cluster. The different scalings for the x- and y-dimension (measured in pixels) to the z-dimension (scales on the image pyramid) must be considered.

Moreover, we developed a minimal density constraint. If one object is tracked in a video stream most particles are contracted near the object. If a second object enters the scene it can take several frames till it is captured by at least one sample. Therefore, we integrated a constraint with a minimal density for each image area (defined by an equidistant grid over the frame and scales of the image pyramid). Within each image area additional samples are randomly distributed until the minimal density constraint is fulfilled.

IV. Active Dual Camera System for HCI Applications and Experiments

We applied the new 3D CCT to an active dual camera system. The system (Fig. 4, *left image*) consists of a large 30" monitor, a static camera (*red box*: Basler A301fc, 8mm lens), a PTZ-camera (*blue*: Sony Evi D100) and two 300W light panels. The *monitor image* of the dual camera system shows a tracking with the samples (*black points*: uncertain and *red*: high certainty). Fig. 4, *right* presents results of the 3D CCT. The distribution of the

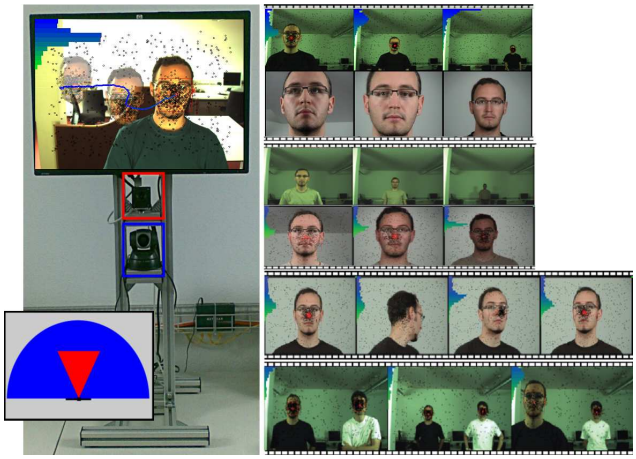

Fig. 4. Our active dual camera system (*left*) demonstrates robust 3D CCT (*right*), even if the object is not visible for some frames (*right*, 5^{th} *row*) and for multiple objects (6^{th} *row*).

samples respective to the density function of the CCT for the third dimension is shown by the histograms in the *upper left corner* of each tracking image. If many samples are distributed on larger scales of the image pyramid (face near to the camera) the maximum of the histogram moves to the *lower (green) bars* and if many samples are on the smaller scales (further away from the camera) it moves to the *upper (blue) bars* (e.g., 1^{st} *row right*). Even for larger distances the PTZ-camera delivers a high resolution image section of the face, making face or expression recognition HCI applications feasible (*right*, 2^{nd} *and* 4^{th} *row*; Note that the maximal optical zoom is already exceed at the 3^{rd} *frame* at this rows). The active dual camera system tracking is more robust to fast movements of the object (CCT on the static camera, 1^{st} *row*, controls the PTZ-camera, 2^{nd} *row*). However, the CCT direct on the PTZ-camera stream (PTZ-camera controls itself, 3^{rd} *row*, and the static camera, 4^{th} *row*, shows only an overview of the scene) can track larger distances and angles because of the larger visible scene area of the PTZ-camera (The *Schema* at Fig. 4, *bottom left* compares the scene area (*red triangle*) of the static camera and the PTZ-camera (*blue*).).

In the experiments we compared the novel CCT with tracking based on Kalman filters [4], on original Condensation [3] and with state-of-art face detection methods [1, 6]. In opposite to the Kalman tracking CCT is able to track multiple objects. Compared to the original Condensation the extended approach can track multiple objects stably over long time periods and on different distances because of the density function on the third dimension (Fig. 4, *right*, 6^{th} *row*). CCT is more robust to temporary occlusion. If objects get lost for some frames the particles distribute faster over the frame and contract again when the object is found back (Fig. 4, *right*, 5^{th} *row*). The introduced 3D CCT yields a more than 10 times faster tracking as state-of-art detection methods.

The HCI application FaceSwap (Fig. 5) demonstrates the high run-time performance and robustness of the 3D CCT. The faces areas are tracked by CCT in three dimensions, cut out and swapped either between persons on the scene or with faces on arbitrary photographs. The demonstration of the CCT is an enjoying eye-capture at presentations and touches questions from the field of perception psychology, e.g., by taking over different identities. The application was inspired by a joint project with the Academy of Art and Design, Basel [20].

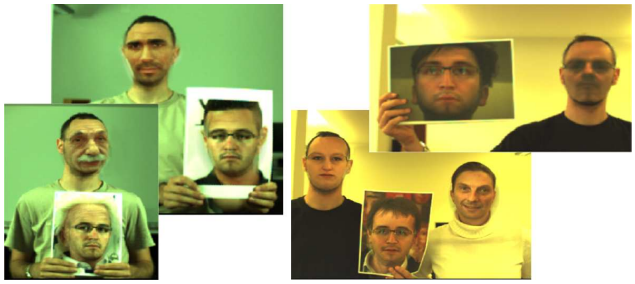

Fig. 5. The HCI application FaceSwap, based on 3D CCT, is an enjoying eye-capture by touching questions from the field of perception psychology.

V. Conclusion

The Condensation and the Wavelet Approximated Reduced Vector Machine approach are joined by the core idea to contract the computational effort to regions (Condensation) or vectors (W-RVM) of the feature space with high statistical likelihood to contain objects of interest. We adapted the W-RVM classifier to tracking, refined the Condensation approach by a coarse-to-fine measurement function and unified both approaches. We applied the Condensation approach for abstract multidimensional feature vectors, e.g. the samples are distributed, based on the now three-dimensional density function, over the x-, y- (in-plane translation) and also the z-dimension (distance) on a camera scene. Moreover, we introduced a robust multi-object tracking by extensions to Condensation. This enables a more natural HCI by tracking a much larger range of distances or tracking different object classes simultaneously in real-time. At the experiments the robustness and efficiency of the novel 3D CCT approach is compared to other tracking and detection methods on an active dual camera system and integrated in HCI applications like FaceSwap touching questions from the field of perception psychology.

Acknowledgment

The authors would like to thank M. Hudritsch's group at the Department of Computer Science, FHNW, Muttenz, in particular, P. Chappuis and D. Blanc for the cooperation at the 'Realtime Face Tracking' project, B. Groß, Academy of Art and Design, Basel, during the 'Attack & Swap' project and J. Batliner, GraVis, University of Basel, during his Bachelor thesis [14].

References

[1] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[2] M. Rätsch, S. Romdhani, and T. Vetter, "Efficient face detection by a cascaded support vector machine using haar-like features," *Proc. DAGM'04: 26th Pattern Recognition Symposium*, pp. 62 – 70, 2004.

[3] M. Isard and A. Blake, "Condensation - conditional density propagation for visual tracking," *IJCV*, 1998.

[4] D. Terzopoulos and R.Szeliski, "Tracking with kalman snakes," *In Active Vision, MIT*, pp. 3 – 20, 1992.

[5] M. Rätsch, S. Romdhani, G. Teschke, and T. Vetter, "Over-complete wavelet approximation of a support vector machine for efficient classification," *Proc. DAGM'05: 27th Pattern Recognition Symposium*, Vienna, 2005.

[6] M. Rätsch, G. Teschke, S. Romdhani, and T. Vetter, "Wavelet frame accelerated reduced support vector machines," *IEEE Transactions on Image Processing*, vol. 17, no. 12, pp. 2456 – 2464, Dec. 2008.

[7] V. Vapnik, *Statistical Learning Theory*. N.Y.: Wiley, 1998.

[8] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *PAMI*, vol. 20, pp. 23–38, 1998.

[9] H. Schneiderman and T. Kanade, "A statistical method for 3d object detection applied to faces and cars," *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 746 – 751, 2000.

[10] H. Kang and D. Kim, "Real-time multiple people tracking using competitive condensation," *Pattern Recognition*, vol. 38, pp. 1045 – 1058, 2005.

[11] L. Bretzner and T. Lindeberg, "Qualitative multiscale feature hierarchies for object tracking," *Journal of Visual Communication and Image Representation*, vol. 11, pp. 115 – 129, 1999.

[12] T. Yang, S. Li, Q. Pan, J. Li, and C. Zhao, "Reliable and fast tracking of faces under varying pose," *7th International Conference on Automatic Face and Gesture Recognition (FGR'06)*, 2006.

[13] F. Huang and T. Chene, "Tracking of multiple faces for human-computer interfaces and virtual environments," *International Conference on Multimedia and Expo (ICME 2000)*, 2000.

[14] J. Batliner, "Multidimensional face tracking using condensation methods," *Bachelor thesis, University of Basel, GraVis*, 2007.

[15] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *ACM SIGGRAPH*, 1999.

[16] S. Prince, J. Elder, Y. Hou, M. Sizinstev, and E. Olevskiy, "Towards face recognition at a distance," *The Institution of Engineering and Technology Conference on Crime and Security*, 2006.

[17] C. Blumer, "Face tracking controlled active camera system," *Master thesis, University of Basel, GraVis*, 2008.

[18] S. Romdhani, P. Torr, B. Schölkopf, and A. Blake, "Efficient face detection by a cascaded support-vector machine expansion," *Proc. The Royal Society A*, November 2004.

[19] J. Platt, "Probabilistic outputs for support vector machines and comparison to regularized likelihood methods," *in Advances in Large Margin Classifier, MA: MIT Press*, 2000.

[20] J. Diessl and B. Groß, "Attack + swap media installation," *Free Frame plug-in for VVVV, http://www.vvvv.org/tiki-index.php*, July 2006.