Proceedings of the 7th IASTED International Conference
Signal Processing, Pattern Recognition and Applications (SPPRA 2010)
February 17 - 19, 2010 Innsbruck, Austria

3D CASCADED CONDENSATION TRACKING FOR MULTIPLE OBJECTS

Matthias Rätsch
University of Basel
Bernoullistrasse 16
CH-4056 Basel, Switzerland
email: matthias.raetsch@unibas.ch

Clemens Blumer
University of Basel
Bernoullistrasse 16
CH-4056 Basel, Switzerland
email: clemens.blumer@unibas.ch

Gerd Teschke
University of Applied Sciences Neubrandenburg
Brodaer Str. 2
D-17033 Neubrandenburg, Germany
email: teschke@hs-nb.de

Thomas Vetter
University of Basel
Bernoullistrasse 16
CH-4056 Basel, Switzerland
email: thomas.vetter@unibas.ch

ABSTRACT
The Condensation and the Wavelet Approximated Reduced Vector Machine (W-RVM) approach are joined by the core idea to spend only as much as necessary effort for easy to discriminate regions (Condensation) and measurement locations (W-RVM) of the feature space, but most for regions and locations with high statistical likelihood to contain the object of interest. We unify both approaches by adapting the W-RVM classifier to tracking and refine the Condensation approach. Additionally, we utilize Condensation for abstract multi-dimensional feature vectors and provide a template based tracking of the three-dimensional camera scene. Moreover, we introduce a robust multi-object tracking by extensions to the Condensation approach. The new 3D Cascaded Condensation Tracking (CCT) for multiple objects yields a more than 10 times faster tracking than state-of-art detection methods. In our experiments we compare different tracking approaches using an active dual camera system for face tracking.

KEY WORDS
Cascaded Condensation Tracking, Wavelet Approximated Reduced Vector Machine, Coarse-to-Fine Particle Filter, Condensation, Multi-Object Face Tracking, Active Dual Camera System

1 Introduction

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-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], [4]. Condensation is able to track objects in a highly cluttered background. The tracking method is a good alternative to the Kalman Filter [5], 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 a template based classifier.

In this paper we propose to combine Condensation tracking with the efficient Wavelet Reduced Vector Machine (W-RVM) [2], [6], [7]. The W-RVM uses a Double Cascade for early rejections of easy to discriminate image locations. The classifier gains a more than 500 fold speedup compared to an original Support Vector Machine [8]. The classifier trains much faster as the Viola and Jones classifier [1] by same detection accuracy and runtime performance and detects about 25 times faster than the Rowley-Baluja-Kanade detector [9] and about $1e^3$ times faster than the Schneiderman-Kanade [10] 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 coarse-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. [11] 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.

A limitation of tracking approaches is also 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. [12] propose a specialized multi-scale tracking like for features different in size or Yang et al. [13] and Huang et al. [14] use specific deformable templates. In contrast, we want to introduce a novel abstract multi-dimensional 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 as in [15]. 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, [16]) 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 optically zoomed. Prince et al. [17] 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 as in [15]. By Yang et al. [13] 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 [8]. 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.

The main contribution of this paper is the unification of the Condensation tracking by Isard and Blake and the double cascaded W-RVM classifier by Rätsch et al. The obtained novel Cascaded Condensation Tracking (CCT) 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. In this paper we will introduce the CCT based on the following core ideas:

- Adaptation of the W-RVM classifier for tracking and providing a probabilistic output (Section 2).
- Condensation for abstract multi-dimensional feature vectors usable for template based tracking instead tracking of object curves. Distribution of the density function and tracking objects over the three dimensions of the camera scene (Section 3).
- Extension of Condensation by a dynamic and adaptive stochastic prediction of the object dynamics (dynamic and adaptive diffusion matrix, Section 3.1).
- Stable multi-object tracking (Section 3.2) by:
    - Adaptive multi-modal probability distribution,
    - Weighted drift function, and
    - Minimal density constraint.

We apply the CCT on an active dual camera system with a still and PTZ camera providing high resolution image sections for Human Computer Interaction (HCI) applications within large camera scenes. In Section 4 we also compare the robustness and run-time performance with state-of-art face detection and tracking approaches.

2 Probabilistic Wavelet Approximated Reduced Vector Machine

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 $\Phi : \mathcal{X} \to F, \;\; \mathbf{x} \mapsto \Phi(\mathbf{x})$. Although F can be high dimensional, it is usually not necessary to explicitly work in that space [8]. 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 a SVM provides a classifier with the largest margin [8], 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, we refer the reader to [2], [6], [7].):

1. **Support Vector Machine:** Use of a SVM [8] classifier that is known to have optimal generalization capabilities.
   (a) SVM: $\Psi_{\mathrm{SVM}} = \sum_{i=1}^{N_x} \alpha_i \, \Phi(\mathbf{x}_i)$, $\mathbf{x}_i$ are the Support Set Vectors (SSV's)
   (b) Decision function:
   $y(\mathbf{x}) \;=\; \mathrm{sgn}\left(\sum_{i=1}^{N_x} \alpha_i \, k(\mathbf{x}, \mathbf{x}_i) + b\right)$ 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: $\Psi_{\mathrm{RVM}} = \sum_{i=1}^{N_z} \beta_i \, \Phi(\mathbf{z}_i)$.
   (b) Decision function:
   $y(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^{N_z} \beta_i \, k(\mathbf{x}, \mathbf{z}_i) + b_{N_z}\right)$.
3. **Double Cascade:** The RVM is reduced in a second step by approximating each RSV by several levels of Wavelet Approximated Reduced Set Vectors (WRSV's) to obtain a Double Cascade. For non-symmetric data (i.e. only few positives to many negatives) an early rejection of easy to discriminate vectors is achieved. It is obtained by the two following cascaded evaluations 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 approximation 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 [6].
5. **Wavelet Frame:** An over-complete wavelet system is used 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 [7], 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.

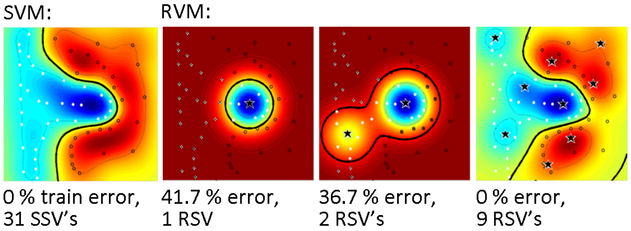

Figure 1: Cascaded application of RSV's (*stars*) to a 2D classification problem (*black and white dots*), showing (*left to right*) the original SVM and the result of using 1, 2, and 9 Reduced Set Vectors.

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.

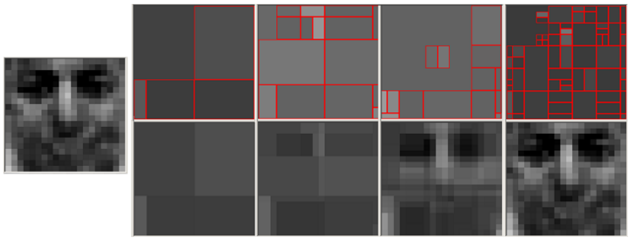

Figure 2: Example of coarse-to-fine W-RSV's for the first RSV (*left*). W-RSV's at different resolution levels (*bottom row*) and the related wavelet approximated residuals (*above*).

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: $(\mathbf{x}_{ffp}|v_{ffp}) = 1/\left(1+\exp\left(c_1\ v_{ffp}+c_2\right)\right)$. The sigmoid fitting trains iterative the parameters c_1 and c_2 of the sigmoid function to map the W-RVM output v_{ffp} of the feature point ffp (e.g. faces or eyes) into probabilities $p_{ffp}\left(\mathbf{x}_{ffp}|v_{ffp}\right)$.

3 3D Cascaded Condensation Tracking for Multiple Objects

3.1 3D Cascaded Condensation Tracking

Condensation, invented by Isard and Blake [3], [4], stands for 'Conditional Density Propagation' and is one of the most successfully used approaches evaluated for different tracking tasks. The main principle of the algorithm is to propagate a density function from one iteration to the next. To this end it uses factored sampling in which the probability distribution of possible interpretations is represented by a randomly generated set. This is called a particle filter, also known as Sequential Monte Carlo methods. The result is highly robust tracking of agile motion. Despite the use of stochastic methods, the algorithm runs in real-time.

Notations: The state of the modeled object at time t is denoted as $\mathbf{x}^{(t)}$. The history of the modeled object at time t is denoted $\mathbf{X}^{(t)} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}\}$. This represents the model feature vector. The state of the observation at time t is denoted $\mathbf{z}^{(t)}$, its history $\mathbf{Z}^{(t)} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(t)}\}$. Further, there is a set of samples $\{\mathbf{s}_1^{(t-1)}, \ldots, \mathbf{s}_n^{(t-1)}\}$ and a set of probabilities $\{\pi_1^{(t-1)}, \ldots, \pi_n^{(t-1)}\}$. Samples are elements of the model feature space which also contain $\mathbf{x}^{(t)}$.

3D Object Tracking: Instead of tracking object curves, the proposed CCT is utilized for template based tracking and can be used for abstract multi-dimensional feature vectors. Therefore, the feature vectors $\mathbf{x}^{(t)}$ and the observation $\mathbf{z}^{(t)}$ can have any dimensions.

In this paper we introduce a tracking of objects within the three-dimensional camera scene. In opposite to other tracking approaches (e.g. [13], [14]) we distribute the samples and track objects not only over the x- and y-coordinates of the image plane, but also over the z-dimension, which is the distance of the camera to image plane (see Fig. 3). Hence, the feature vector $\mathbf{x}^{(t)}$ is three-dimensional ($\mathbf{x}^{(t)}, \mathbf{s}_i^{(t)} \in \mathbb{R}^3$). Similar to conventional object detection approaches [9], an image pyramid of the frame is used in order to locate objects of different sizes and the distance to the camera is represented by the scale of the image pyramid. The observation $\mathbf{z}^{(t)}$ represents the image features from a section of a video frame (e.g. a 20×20 grey value patch, $\mathbf{z}^{(t)} \in \mathbb{R}^{400}$) modeled by the center point $\mathbf{x}^{(t)}$.

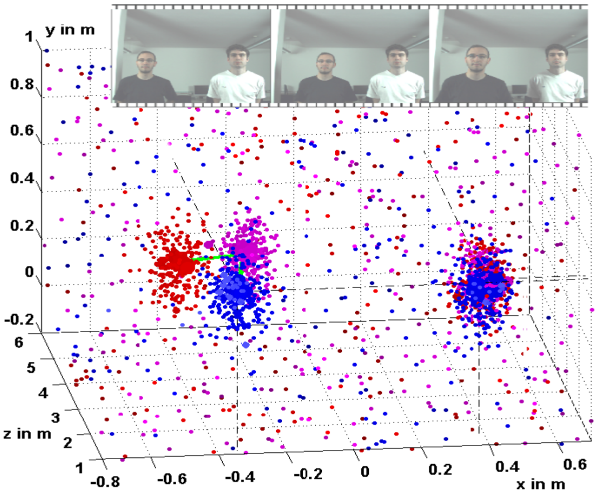

Figure 3: CCT enables tracking of multiple hypotheses in tree dimensions (*chain dotted lines* indicate the end positions). *Red* samples belong to the 1st, *pink* to the 2nd and *blue* to the last frame (*top row, left to right*). The *green line* indicates the track of the left person from left to right and then to the front. The sizes of the samples indicate the weight per sample. The experiment demonstrates the dynamic probability distribution over tree dimensions.

Assumptions: The detection of the likelihood of the object's position within the model $p\left(\mathbf{x}^{(t)}|\mathbf{z}^{(t)}\right)$, given the image signal information at time step t, is not trivial. Therefore, the Bayesian theorem is applied to simplify computation:

$$\begin{aligned} p\left(\mathbf{x}^{(t)}|\mathbf{z}^{(t)}\right) &= \frac{p\left(\mathbf{z}^{(t)}|\mathbf{x}^{(t)}\right)p\left(\mathbf{x}^{(t)}\right)}{p\left(\mathbf{z}^{(t)}\right)} \\ &= kp\left(\mathbf{z}^{(t)}|\mathbf{x}^{(t)}\right)p\left(\mathbf{x}^{(t)}\right) \qquad (1) \end{aligned}$$

The quotient $1/p\left(\mathbf{z}^{(t)}\right)$ is independent of $\mathbf{x}^{(t)}$ and can be expressed by a constant term k. Evaluation of $p\left(\mathbf{z}^{(t)}|\mathbf{x}^{(t)}\right)$ instead of $p\left(\mathbf{x}^{(t)}|\mathbf{z}^{(t)}\right)$ is one of the central concepts of the Condensation approach. It tries to estimate the probability density function for areas of the image with high a-priori likelihood. The prior obtained from the last frame is used to control the density of the samples over the model space. At the sampled feature points of the model the likelihood is measured anew.

Furthermore, we assume for this problem that all observations during the process are independent from each other. This means that:

$$p\left(\mathbf{Z}^{(t)}|\mathbf{X}^{(t)}\right) = \prod_{i=1}^{t} p\left(\mathbf{z}^{(i)}|\mathbf{x}^{(i)}\right) \qquad (2)$$

The second assumption states that the process is a Markov chain, i.e. observations are independent of earlier states:

$$p\left(\mathbf{x}^{(t)}|\mathbf{X}^{(t-1)}\right) = p\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right) \qquad (3)$$

This expresses that observations are only dependent on the last state.

Initialization: For initialization the samples are distributed in the image feature space, which means scattering them over the frame and, because we aspire a three-dimensional object tracking and density function, additional over the scales of the image pyramid. This can be done e.g. randomly or aligned in a grid. For this experiment, we decided to scatter the samples according to a two-dimensional normal distribution in the x, y-plane and uniformly in the scales. All samples are assigned with the same probability of $1/n$.

Selection: Factored sampling is utilized in this step to select the samples that are used for one iteration loop. The probabilities of the samples sum up to one. We can assign a subinterval to every sample in $[0, 1]$ such that the length of the interval is equal to the probability. We now generate a random number r between zero and one and select the sample in whose subinterval the number is situated. Let's say the random number is within the j'th subinterval (Fig. 4). We therefore choose the sample $\mathbf{s}_j^{(t-1)}$ and set $\tilde{\mathbf{s}}_i^{(t)} = \mathbf{s}_j^{(t-1)}$. This is repeated until all n samples are chosen.

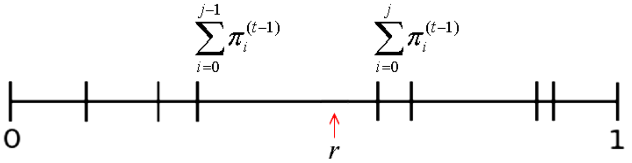
$\sum_{i=0}^{j-1}\pi_i^{(t-1)}$
$\sum_{i=0}^{j}\pi_i^{(t-1)}$
0
r
1

Figure 4: Selection of a sample.

Dynamic Adaptive Prediction: In this step, we want to predict the new position of the samples. Prediction means sampling from $p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)} = \tilde{\mathbf{s}}_i^{(t)})$ to choose each $\mathbf{s}_i^{(t)}$. It is attempted to predict the object's position $\mathbf{x}^{(t)}$ given that the model of the last step was at position $\tilde{\mathbf{s}}_i^{(t)}$. If the dynamics are modeled as a linear stochastic process, we can compute new samples in the following way:

$$\mathbf{s}_i^{(t)} = \mathbf{A}\tilde{\mathbf{s}}_i^{(t)} + \mathbf{B}\mathbf{w}_i^{(t)} \qquad (4)$$

A deterministic and a stochastic component is used, where $\mathbf{A}$ is a translation matrix (drift due to the deterministic component of object dynamics), $\mathbf{B}$ a diffusion matrix and $\mathbf{w}^{(t)}$ a vector of standard normal variates (random component of object dynamics). Matrix $\mathbf{A}$ accounts for the movement of the samples and is detailed in Section 3.2. The matrix $\mathbf{B}$ can be learned a-priori and used constantly [3], [4], [20].

One goal of the proposed work was to find and compare alternatives to control the diffusion matrix $\mathbf{B}$. We realized a constant diffusion matrix, a dynamic diffusion matrix, and a dynamic and adaptive approach. A dynamic diffusion matrix can be computed for each frame from the covariance matrix, i.e. with $\mathbf{B} = 1/\left(n-1\right)\mathbf{D}\mathbf{D}^T$, where $\mathbf{D}$ is the mean-free data matrix constructed from the n samples. As novelty we introduce a dynamic and adaptive approach to compute the diffusion, namely

$$\mathbf{s}_i^{(t)} = \mathbf{A}\tilde{\mathbf{s}}_i^{(t)} + C\left(1 - \pi_i^{(t)}\right)\mathbf{B}\mathbf{w}_i^{(t)} \qquad (5)$$

where C is a constant that represents the scatter. The approach is adaptive because it diffuses samples with low weight more than samples with high weight and is dynamic because the diffusion is new adapted at each time step t. This extension increases the localization accuracy of the tracked object (because on samples $\tilde{\mathbf{s}}_i^{(t)}$ with higher weight $\pi_i^{(t)}$ less noise is added) by no additional computational effort. A smaller density is needed for background image areas, because on samples $\tilde{\mathbf{s}}_i^{(t)}$ with smaller weight $\pi_i^{(t)}$ more noise is added. Entering objects into the camera scene or lost objects are faster detected by fewer samples at these feature space areas.

The dynamic adaptive diffusion matrix enables a higher accuracy of the tracking locations by no increase of complexity. The scatter parameter C controls the trade-off between the robustness of the tracking on one hand and a complexity reduction and an increase of the localization accuracy on the other hand. Moreover, the multi-modality of the density function can be controlled by the dynamic adaptive diffusion. However, for a stable multi-object tracking over a longer time period more extensions are necessary and introduced in Section 3.2.

Measurement: In this step, the samples are measured and their probabilities are updated. Now that the samples are placed in the area where the object is presumed to be, they are measured in terms of $\mathbf{z}^{(t)}$: $\pi_i^{(t)} = p(\mathbf{z}^{(t)}|\mathbf{x}^{(t)} = \mathbf{s}_i^{(t)})$ which means that we assign to $\pi_i^{(t)}$ the likelihood that the object is observed in the image at the position $\mathbf{x}^{(t)}$ of the model, represented by the drifted and diffused feature vector $\mathbf{s}_i^{(t)}$. Condensation uses statistics to distribute the samples $\mathbf{s}_i^{(t)}, i = 1, \dots, n$ by a conditional probabilistic density function over the feature model space (e.g. an image pyramid) and measures only at this certain pixels of the image if an object of interest is located at these image positions. Instead of all pixels, used for object detection, a much lower number n of measurements is needed. This provides a significant speedup. The W-RVM approach joins the same concept only to spend as many operations as necessary to easy to discriminate model space regions, but most for locations where objects of interest are predicted by statistical assumptions. As novelty we combine both approaches for a reduction of computational complexity by refining the measurement function. Instead using a constant number of operations, as used in former Condensations methods, we adapted and integrated the W-RVM as measurement function. The W-RVM uses a Double Cascade and other methods to contract computational complexity only to vectors with higher statistical interest as summarized in Section 2.

The proposed Cascaded Condensation Tracking yields an optimal contraction of computational complexity per region (based on Condensation) and per vector (based on W-RVM) of the feature space. This twice stochastically contracted complexity per region (symbolized by Voronoi-diagram areas) is demonstrated on an example image in Fig. 5 where for difficult to discriminate feature space regions (*pink*) more operations per vector (W-RVM) and a higher sample density (Condensation) are used, than for homogenous background (*white*). The complexity per region is colorized as number of operations (used by W-RVM) per sample and dived by the size of the Voronoi area (symbolizing the density function obtained by Condensation).

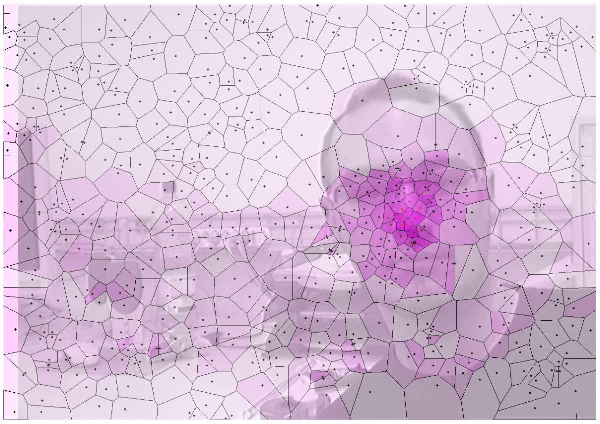

Figure 5: The computational complexity is twice contracted to regions with high probability to contain objects of interest. For difficult to discriminate feature space regions (*pink*) more operations per measurement location (W-RVM) and a higher sample density (Condensation) are used, than for homogenous areas (*white*).

Object Position: The position of the object can be estimated using the following formula for the expectation (the object location estimation is detailed for multi-objects in Section 3.2):

$$E\left[\mathbf{x}^{(t)}\right] = \sum_{i=1}^{n} \pi_i^{(t)} \mathbf{s}_i^{(t)} \tag{6}$$

The CCT performs one loop per time step (frame) consisting of selection, prediction and measurement. Samples are selected, then drifted and diffused. Finally, the new weights are measured, the next iteration can start.

3.2 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 the original Condensation algorithm is that a multi-object tracking is not stable over longer time periods, although it provides a multi-modal density function and probability distribution (function with more than one maximum) as opposite to the Kalman Filter [5]. For the maxima at the density function we use the same clustering approach as in [7], but here by assigning samples to clusters with respect to their weight and their Euclidian distance to the cluster centers. The object positions (cluster centers $\mathbf{c}$) are estimated by (6) over the assigned samples to each cluster.

In the original Condensation algorithm the cluster with a higher probability to be an object of interest draws off samples from improbably clusters (see Fig. 7, *top*). The not as probably cluster 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 [11], 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.

The novel *adaptive multi-modal probability distribution* uses one multi-modal distributed set of samples but adapt the probability distribution individual for every cluster. The original probability distribution is manipulated to suppress samples of other clusters (Fig. 7, *bottom* shows stable multi-object tracking). No expensive computations are needed. From the probability distribution vector a manipulated probability distribution matrix with size $n \times m$ is calculated where n is the number of samples and m the number of clusters. The manipulated likelihoods $\pi_{i,j}$:

$$\pi_{i,j} = \pi_i \prod_{k=0,k\neq j}^{m-1} \left(1 - \frac{1}{\exp\left(\left(\frac{d_{i,k}}{p}\right)^q\right)}\right) \tag{7}$$

where $d_{i,k}$ is a distance measurement and p, q are empirical constants (we obtained good results with $p = 40$ and $q = 6$). The new $\pi_{i,j}$ are normalized.

The selection process is also adapted so that n/m-times every column of the manipulated probability distribution matrix is used. This balances the amount of samples per cluster. A stable tracking of multiple objects is obtained over long time periods.

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.

The drift in Blake's approach is calculated by a stochastic differential equation for single movements [20]. 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 multiple objects can move in different directions and with different speed. We obtain a weighted deterministic component of object dynamics in (4) by defining the translation matrix $\mathbf{A}$ by:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & f_x \\ 0 & 1 & 0 & f_y \\ 0 & 0 & 1 & f_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{8}$$

For $\mathbf{A}$ and Equation (4) and (5) homogeneous coordinates are used. The translation vector $\mathbf{f}$ is defined by:

$$\mathbf{f} = \sum_{i=1}^{m} \left(1 - \frac{1}{\sum_{j=1}^{m} \Delta\mathbf{c}_j^{(t-1)}} \left(\mathbf{s} - \mathbf{c}_i^{(t-1)}\right)\right) \Delta\mathbf{c}_i^{(t-1)} \tag{9}$$

The cluster offsets are described by $\Delta\mathbf{c}^{(t)} = \mathbf{c}^{(t)} - \mathbf{c}^{(t-1)}$. The weights are evaluated by the component wise distance to the cluster centers and normalized. The drift of a sample is continually most influenced by the drift of the nearest cluster.

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.

4 Active Dual Camera System for CCT Experiments and Results

We applied the new 3D CCT to an active dual camera system. The system (Fig. 6, *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. Fig. 6, *right* demonstrates results of the 3D CCT. The distribution of the 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* (*right, 1st row*). Even for larger distances the PTZ-camera delivers a high resolution image section of the face, making face or expression recognition HCI applications feasible (*2nd and 4th row*, Note that the max. optical zoom is already exceed at the 3rd frame). The active dual camera system tracking is more robust to fast movements of the object (CCT on the static camera, *1st row*, controls the PTZ-camera, *2nd row*).

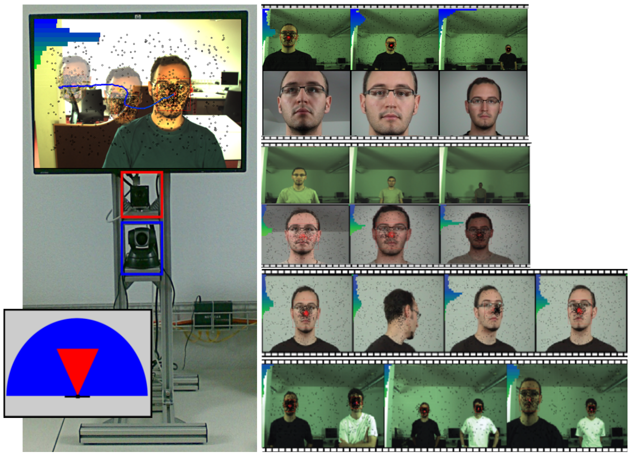

Figure 6: Our active dual camera system (*left*) demonstrates robust CCT in three dimensions (*right*), even for short occlusion of the object and for multiple objects (*right, 5th and 6th row*).

However, the CCT direct on the PTZ-camera stream (PTZ-camera controls itself, *3rd row*, and the static camera, *4th 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. 6, *bottom left* compares the scene area (*red triangle*) of the static camera and the scene area (*blue*) of the PTZ-camera.).

To compare different approaches in the experiments we used a video sequence of 1000 frames collected with the dual camera system. On each frame the faces of two persons are visible. We compared the novel CCT with tracking based on Kalman filters [5], with the original Condensation [4] and with state-of-art face detection methods [1], [7].

In opposite to the Kalman tracking Condensation is able to track multiple objects. On the test set in 228 of 1000 frames both faces where correctly detected by the original Condensation. In comparison the CCT tracked 994 frames correctly. This experiment demonstrates that the CCT can track multiple objects stably over long time periods. Because of the density function on the third dimension the tracking is also stable on different distances (Fig. 6, *right, 6th row* shows examples of the test set). Compared to original Condensation, CCT was more robust to temporary occlusion at the experiments. If objects get lost for some frames (e.g. Fig. 6, *right, 5th row*), the particles distribute faster over the frame and contract again when the object is found back, because of dynamic adaptive diffusion matrix.

Fig. 7 shows on a subset of the test set (see example frames in Fig. 6, *right, 6th row*) that the CCT can stably track two persons, taking advantage of the adaptive multimodal probability distribution. The density function is projected onto the horizontal translation axis.

We also compared the run-time performance at the experiments on a standard PC (Dual-Core, 2.3GHz) for detection on an equidistant grid and for the Condensation tracking (both using different number of scales, therefore the time is given per scale of the image pyramid). As classifier either a standard SVM [8] or the W-RVM (training and data as described in [7]) is used with comparable detection accuracy. Table 1 shows that the contraction of the computational complexity either per region (based on Condensation) or per location (based on the double cascaded W-RVM) of the feature space improves the run-time significantly. However, best performance, by no significant loss of accuracy, is gained joining both contractions of computational complexity over the feature space by the Cascaded Condensation Tracking. The introduced 3D CCT yields a more than 10 times faster tracking as state-of-art detection methods.

Approach	sec per scale
Detection with SVM classifier	6.76
Detection with double cascaded W-RVM	0.0135
Condensation with SVM classifier	0.329
Condensation with W-RVM (CCT)	0.00133

Table 1: Comparison of run-time performance.

5 Conclusion

The Condensation and the Wavelet Approximated Reduced Vector Machine approach 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

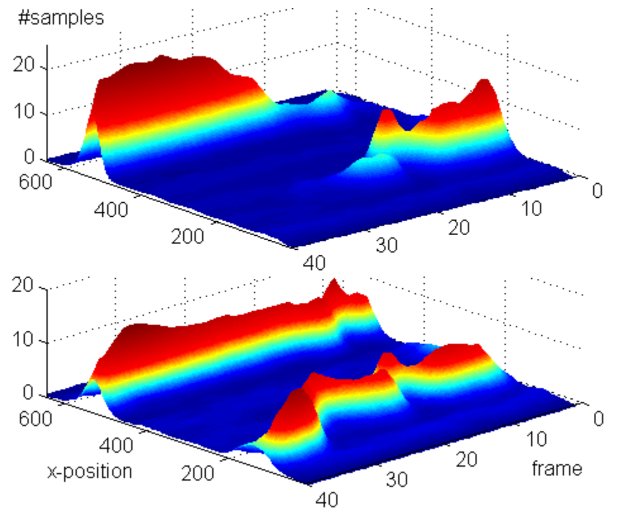

Figure 7: The projection onto the horizontal translation axis of the density function demonstrates that after some iterations one cluster can dominate by taking over all samples (*top*, original Condensation). Based on the novel adaptive multi-modal probability distribution the multi-object CCT is stable (*bottom*).

the feature space, but most for locations with high statistical likelihood to contain the object of interest. In this paper both approaches are unified. We adapted the W-RVM classifier to tracking (e.g., the W-RVM provides now a probabilistic output) and refined the Condensation approach by a Double Cascade measurement function. Additionally, we generalized the Condensation approach for abstract multi-dimensional 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 like the adaptive probability distribution or the minimal density constraint. The robustness and efficiency of the 3D CCT approach is demonstrated on an active dual camera system and compared with other approaches. The introduced 3D Cascaded Condensation Tracking for multiple objects yields a more than 10 times faster tracking as state-of-art detection methods. This enables more natural HCI by tracking a much larger range of distances or tracking different object classes (like faces, eyes, and mouth corners) simultaneously in real-time.

Acknowledgements

This work was supported by a grant from the German Research Foundation (DFG TE 354/4-1). The authors would like to thank M. Hudritsch's group at the Department of Computer Science, FHNW, Muttenz, Basel, and J. Batliner, GraVis, University of Basel for their work and cooperation in preceding projects.

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*, pages 62 – 70, 2004.

[3] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In *Proc. European Conference Computer Vision*, pages 343 – 356, Cambridge, 1996.

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

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

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

[7] M. Rätsch, G. Teschke, S. Romdhani, and T. Vetter. Wavelet frame accelerated reduced support vector machines. *IEEE Transactions on Image Processing*, 17(12):2456 – 2464, Dec. 2008.

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

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

[10] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to face and cars. In *CVPR*, pages 746–751, 2000.

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

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

[13] 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.

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

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

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

[17] 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.

[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*, 460(2501):3283 – 3297, 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] A. Blake, M. Isard, and D. Reynard. Learning to track the visual motion of contours. *Artificial Intelligence*, 78:179 – 212, 1995.