3D Cascaded Condensation Tracking for Multiple Objects

Submitted for Blind Review

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) or vectors (W-RVM) of the feature space, but most for regions with high statistical likelihood to contain the object of interest. In this paper we unify both approaches by adapting the W-RVM classifier to tracking and refine the Condensation approach. Additionally, we utilize Condensation for abstract multidimensional 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. We demonstrate more natural HCI applications by tracking a large camera scene with an active dual camera system.

1. 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 with it. A machine which can detect objects is since many years an important aspect of computer science. Especially that a machine can localize a human's 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], [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 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 speed-up compared to an original Sup-

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. [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.

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. [12] propose a specialized multiscale 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 optical 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 [18]. 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 3.1).
- Extension of Condensation by a dynamic and adaptive stochastical prediction of the object dynamics (dynamic and adaptive diffusion matrix, Sec-
2. 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 double cascaded W-RVM measurement function.

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 labelled training set consisting of a series of e.g. 20 × 20 image patches \( x_i \in X \) (arranged in a 400 dimensional vector) along with their class labels \( y_i \in \{-1, 1\} \). Support Vector classifier implicitly map the data \( x_i \) into a dot product space \( F \) via a (usually nonlinear) map \( \Phi : x \rightarrow F, x \mapsto \Phi(x) \). 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 \( k(x, x') \) to compute the dot products in associated feature spaces, i.e. \( k(x, x') = \langle \Phi(x), \Phi(x') \rangle \). The training of an 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 an SVM [8] classifier that is known to have optimal generalization capabilities.
   - SVM: \( \Psi_{svm} = \sum_{i=1}^{N_x} \alpha_i \Phi(x_i), x_i \) are the Support Vectors (SSV’s)
   - Decision function: \( y(x) = \text{sgn}\left(\sum_{i=1}^{N_x} \alpha_i k(x, x_i) + b\right) \)
     with kernel function \( k(\cdot, \cdot) \), e.g. Gaussian kernel \( k(x, z) = \exp(-\|x-z\|^2/(2\sigma^2)) \).

2. **Reduced Support Vector Machine:** The SVM is reduced by an RVM [19]. The RVM uses a reduced set of Support Vectors. Figure 1 shows that with less Reduced Set Vectors (\( N_y < N_x \)) the same decision accuracy can be obtained. In contrast to the SSV’s, the RSV’s are not a subset of the training set.
   - RVM: \( \Psi_{rvm} = \sum_{i=1}^{N_y} \beta_i \Phi(z_i), z_i \) are the Reduced Set Vectors (RSV’s)
   - Decision function: \( y(x) = \text{sgn}\left(\sum_{i=1}^{N_y} \beta_i k(x, z_i) + b\right) \).

3. **Double Cascade:** The RVM is approximated in a second step by approximating each RSV by several levels of Wavelet Approximated Reduced Set Vectors (W-RSV’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:
   - **Cascade over the number of used W-RSV’s** (see Figure 1) and
   - **Cascade over the resolution levels of each W-RSV** (see Figure 2).
   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 overcomplete 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. Example demonstrating the cascaded RVM The result of the 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. Darker regions indicate strong support (higher certainty) for the classification. Romdhani et al. [19] showed that toy example that with only 9 RSV’s instead of 31 Support Set Vectors the RVM gains the same error rate as the SVM. Using only the first RSV’s yields high error rates, but data points (with a large negative distance to the classification boundary) can be early rejected as negative points, without further evaluation cost.

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 a cascaded W-RSV for the 1\(^{\text{st}}\) RSV Left: first RSV; Right: W-RSV’s at different resolution levels (bottom row) and the related wavelet approximated residuals (above). Already on the first approximations levels (left to right, stages of the 2\(^{\text{nd}}\) cascade) first image locations (like homogenous background) can be rejected. Only for more difficult image locations the full resolution (complexity) of the W-RSV’s must be used (bottom right).

The sigmoid function fitting is a model-trust algorithm, based on the Levenberg-Marquardt algorithm [20]. 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 regularised binomial maximum likelihood. A two-parameter sigmoid is chosen as the post-processing, since it matches the posterior that is empirically observed.

\[
p_{f_{SV}}(x_{f_{SV}} | v_{f_{SV}}) = \frac{1}{1 + \exp(c_1 v_{f_{SV}} + c_2)}
\]

The sigmoid fitting trains iterative the parameters \(c_1\) and \(c_2\) of the sigmoid function to map the W-RVM output \(v_{f_{SV}}\) of the feature point \(f_{SV}\) (e.g. faces or eyes) into probabilities \(P_{f_{SV}}(x_{f_{SV}} | v_{f_{SV}})\).

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 \(x(t)\). The history of the modeled object at time \(t\) is denoted \(X(t) = \{x(1), \ldots, x(t)\}\). This represents the model feature vector. The state of the observation at time \(t\) is denoted \(z(t)\), its history \(Z(t) = \{z(1), \ldots, z(t)\}\). Further, there is a set of samples \(\{s_1^{(t-1)}, \ldots, s_n^{(t-1)}\}\) and a set of probabilities \(\{p_1^{(t-1)}, \ldots, p_n^{(t-1)}\}\). Samples are elements of the model feature space which also contains \(x(t)\).

3D Object Tracking: Instead of tracking object curves, the proposed Cascaded Condensation Tracking (CCT) is utilized for template based tracking and can be used for abstract multi-dimensional feature vectors. Therefore, the feature vectors \(x(t)\) and the observation \(z(t)\) 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 Figure 3). Hence, the feature vector \(x(t)\) is three-dimensional \((x(t), s(t) \in \mathbb{R}^3)\). 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 $z^{(t)}$ represents the image features from a section of a video frame (e.g. a $20 \times 20$ grey value patch, $z^{(t)} \in \mathbb{R}^{400}$) modeled by the centre point $x^{(t)}$.

The probability of the history of the image features given the history of the modeled object is the product of the single probabilities.

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

$$p(x^{(t)} | x^{(t-1)}) = p(x^{(t)} | x^{(t-1)}) \quad (4)$$

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 2-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 (Figure 4).

$$\sum_{i=0}^{j-1} \pi_{ij}^{(t-1)} = \sum_{i=0}^{j} \pi_{ij}^{(t-1)}$$

Figure 4. Selection of a sample

We therefore choose the sample $s_{i}^{(t-1)}$ and set $\tilde{s}_{i}^{(t)} = s_{i}^{(t-1)}$. This is repeated until all $n$ samples are chosen.

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

$$s_{i}^{(t)} = A \tilde{s}_{i}^{(t)} + B w_{i}^{(t)} \quad (5)$$

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 [4], [21], [22].

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
$\mathbf{B} = 1/(n-1)\mathbf{D}\mathbf{D}^T$, 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

$$ s^{(t)}_i = \mathbf{A}s^{(t)}_i + \mathbf{C}(1 - \pi^{(t)}_i)\mathbf{B}w^{(t)}_i, \quad (6) $$

where $\mathbf{C}$ 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 $s^{(t)}_i$
with higher weight $\pi^{(t)}_i$ less noise is added) by no addi-
tional computational effort. A smaller density is
needed for background image areas, because on samples $s^{(t)}_i$
with smaller weight $\pi^{(t)}_i$ 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 in-
crease of complexity. The scatter parameter $\mathbf{C}$ 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. More-
ever, the multi-modality of the density function can be controlled by the dynamic adaptive diffusion. How-
ever, for a stable multi-object tracking over a longer time period more extensions are necessary and intro-
duced 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 $x^{(t)}$:

$$ \pi^{(t)}_i = p\left(z^{(t)} | x^{(t)} = s^{(t)}_i \right) \quad (7) $$

which means that we assign to $\pi^{(t)}_i$ the likelihood that the object is observed in the image at the position $x^{(t)}$ of the model, represented by the drifted and diffused feature vector $s^{(t)}_i$ of the new selected sample $s^{(t)}_i$. Condensation uses statistics to distribute the samples $s^{(t)}_i, i = 1, \ldots, n$ 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 dis-
riminate 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 refin-
ing 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 inter-
est 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 Figure 5.

**Figure 5. Twice reduced complexity per region**

The computational complexity is twice contracted to regions (here in 2D symbolized by Voronoi-diagram areas around the sample locations) with high probability to contain the object of interest. 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 (symboliz-
ing the density function obtained by Condensation).

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

\[ E[x^{(i)}] = \sum_{i=1}^{n} p_i^{(i)} s_i^{(i)} \]  

(8)

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, \(c\)) are estimated by (8) over the assigned samples to each cluster.

Figure 6 shows an example for a one dimensional density function (left) and probability distribution (right) with two clusters. Figure 6, left illustrates the effect in the original Condensation algorithm that after some iterations one cluster dominates. Condensation contracts more and more samples around the cluster with a higher probability to be an object of interest and draws off samples from improbably clusters. 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.

Our 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 as Figure 6, right shows. 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 \(\pi_{i,j}\) are:

\[ \pi_{i,j} = \pi_i \prod_{k=0,k\neq j}^{m-1} \left(1 - \frac{1}{\left(\frac{d_{i,k}}{e^p}\right)^q}\right) \]  

(9)

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.

![Figure 6. Multi-object CCT with a single set of particles using an adaptive multi-modal probability distribution](image-url)

**Left column:** By the original Condensation the most probably cluster (cluster 1) obtains in every iteration (T0 till T3) more and more samples and the sample density of cluster 2 is reduced to the background density. **Right column:** The original probability distribution with two clusters (top) is two times adaptively manipulated, one with respect to cluster 1 (samples near cluster 2 suppressed, middle) and to cluster 2 (bottom).

The number of objects can be limited (e.g., if only one person is in the image) to \(c_{\text{max}}\) clusters. To profit from a-priori knowledge the multi-object certainties are calculated for all found clusters and the best \(c_{\text{max}}\) clusters are kept. After calculating the weighted certainties the dispensable cluster regions (clusters not in \(c_{\text{max}}\)) 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 [22]. 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 (5) by
defining the translation matrix $\mathbf{A}$ by:

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

(10)

For $\mathbf{A}$ and Equation (5) and (6) homogeneous coordinates are used. The translation vector $\mathbf{f}$ is defined by:

$$
\mathbf{f} = \sum_{j=1}^{c} \left( 1 - \frac{1}{\sum_{j=1}^{c} \Delta \mathbf{c}_{j-1}} \right) \Delta \mathbf{c}_{j-1}^{-1} \mathbf{s} - \mathbf{c}_{j-1}
$$

(11)

The cluster offsets are described by $\Delta \mathbf{c} = \mathbf{c} - \mathbf{c}^t$. 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 HCI applications and results

We applied the new 3D CCT to an active dual camera system. The system (Figure 7, 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. Figure 7, 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 (e.g., 1st 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, 2nd and 4th row). 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 7. Results of CCT using a dual camera system

Left: The monitor image of the dual camera system shows a tracking with samples (black points: uncertain and red: high certainty), the distribution of the samples over the scales of the image pyramid (blue to green histogram), and the track of the CCT as blue line. Right: 3D CCT on the static camera (1st row) controls the PTZ-camera (2nd row); 3D CCT performed directly on the PTZ-camera (4th row). Note that the max. optical zoom is already exceeded at the 3rd frame in the 2nd and 4th row. CCT is robust, even if the object is not visible for some frames (5th row) and multiple objects can be tracked (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 Figure 7, bottom left compares the scene area (red triangle) of the static camera and the scene area (blue) of the PTZ-camera.). CCT is robust to temporary occlusion. If the object gets lost for some frames the particles distribute over the frame and contract again when the object is found back (right, 5th row). CCT is able to track multiple objects, even on different distances because of the density function on the third dimension (right, 6th row).

The HCI application FaceSwap (Figure 8) 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 [23].
and the from

Figure 8. HCI application FaceSwap applying 3D CCT
FaceSwap is an enjoying eye-capture by touching questions from the field of perception psychology. The performance of the 3D CCT is demonstrated by tracking multiple persons and swapping their faces or with faces on photographs.

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 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 introduced 3D Cascaded Condensation Tracking for multiple objects yields a more than 10 times faster tracking as state-of-art detection methods [7]. This enables a 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. The robustness and efficiency of the 3D CCT approach is demonstrated on an active dual camera system for HCI applications like FaceSwap.

6. References


