Fast and Adaptive Deep Fusion Learning for Detecting Visual Objects

Nikolaos Doulamis[1] and Anastasios Doulamis[2]

[1] National Technical University of Athens,
9 Heroon Polytechnious Str. 15773, Zografou, Athens, Greece
[2] Technical University of Crete, University Campus, Chania, Greece.
{ndoulam,adoulam}@cs.ntua.gr

Abstract. Currently, object tracking/detection is based on a "shallow learning" paradigm; they locally process features to build an object model and then they apply adaptive methodologies to estimate model parameters. However, such an approach presents the drawback of losing the "whole picture information" required to maintain a stable tracking for long time and high visual changes. To overcome these obstacles, we need a "deep" information fusion framework. Deep learning is a new emerging research area that simulates the efficiency and robustness by which the humans' brain represents information; it deeply propagates data into complex hierarchies. However, implementing a deep fusion learning paradigm in a machine presents research challenges mainly due to the highly non-linear structures involved and the "curse of dimensionality". Another difficulty which is critical in computer vision applications is that learning should be self adapted to guarantee stable object detection over long time spans. In this paper, we propose a novel fast (in real-time) and adaptive information fusion strategy that exploits the deep learning paradigm. The proposed framework integrates optimization strategies able to update in real-time the non-linear model parameters according in a way to trust, as much as possible, the current changes of the environment, while providing a minimal degradation of the previous gained experience.

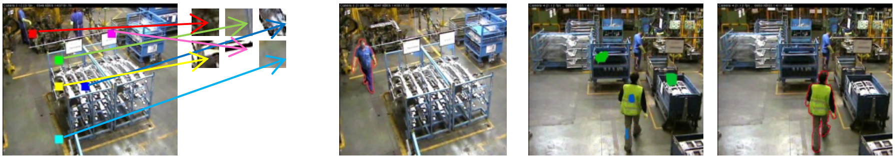
1 Introduction

The current object detection methods exploit a "shallow learning paradigm"; they locally process and map features to build an object model and then they apply adaptive learning methods to estimate model parameters [1]. However, the use of local features inherently presents the drawback of losing the "whole picture information" resulting in several mismatches (see Fig. 1a). Although attempts have been recently proposed to solve this critical aspect through global optimization strategies [2], it seems that there is no a unified mathematical framework that allows "deep" information fusion under a fast (in real-time) and adaptive way (robust to environmental visual changes). The current approaches present the drawback that the training procedures are highly unstable; we need several implementation cycles to conclude to a stable solution.

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part III, LNCS 7585, pp. 345–354, 2012.

However, the main difficulty in implementing such a deep information fusion algorithm is the so called "curse of dimensionality", i.e., the learning complexity exponentially grows with a linear increase in the dimensionality of the data. To make things worse, the data should be related with highly non-linear associations, which present additional research challenges to be optimized and adapted under real-time constraints. This is the reason why until now only shallow learning paradigms have been adopted in computer vision.

Although multi-layer learning models have been known for many years ago, they could not been trained well due to the fact that the performance of the existing training algorithms is significantly deteriorated for large number of hidden layers. This drawback was alleviated, to an extent, when a reasonably efficient, new learning algorithm was introduced by Hinton et al. [3], opening new frontiers for the use of deep structures. However, even with these significant contributions, deep information fusion learning lacks self adaptability which is a critical aspect in computer vision; object tracking for very long time periods encounters abrupt and high visual changes (see Fig.1(b)-(d)). For this reason, semi-supervised learning strategies (SSL) have been investigated as an efficient learning paradigm to increase the reliability of object tracking [1] using, however, shallow boosting mechanisms. However, again semi-supervision (e.g., simple inclusion of unlabelled) does not face the inherent problem of instability and time consuming training process which is presented in deep, non-linear structures.



(a) (b) (c) (d)

Fig. 1. (a) The deep learning paradigm; local processing looses the 'whole picture information. Although one could discriminate the content of the whole image, it is impossible to understand where the six sub-images come from. (b,c) The self training necessity; It is impossible for the tracker to remain stable for both images due to background/foreground changes. Thus, we need self-training mechanisms; For example, the green (blue) regions are selected as confident background (foreground) from the pool of unlabelled data by exploiting motion information (see Fig.1c). (d) The final tracking after several adaptation cycles.

1.1 Previous Works

Recently a great effort has been dedicated to handle object tracking as a classification problem [4]. However, these approaches exploit no adaptable mechanisms to update the performance of the classifier and thus the structure of the model remains fixed. One of the first approaches towards an adaptable classification for object tracking have been presented in [5], [6]. These works, however, do not face efficiently the general trade-off between model stability and adaptability. A highly specific model significantly increases the reliability of the tracker to capture the target but looses adaptability. On the other hand, a highly general model copes with the legitimate changes in appearance but with a cost in reliability. To cope with these problems, Matthews et al. [7] propose an updated template algorithm that avoids the "drifting" inherent. Other methods exploit the semi-supervised paradigm [1], [8] a co-training strategy [9], a combination of generative and discriminative trackers [10] or finally coupled layered visual models [11].

The realization that local information sometimes is not enough to make a correct decision led to the development of global optimization trackers. These trackers operate on larger time scales and make use of high-level reasoning to re-solve ambiguities. Examples are the use of a minimum cost graph matching that runs the Hungarian algorithm [12]. Other works handle the problem as a minimum flow cost problem [13], [2]. However, in case that the background / foreground significantly change, there is no way to estimate matches for long time periods forcing the algorithm to fail.

1.2 Contribution

In this paper, we propose a novel mathematical framework which permits fast and adaptive information fusion over deep learning structures. Very few works have been proposed in the literature that exploit conventional (i.e., non-adaptive) deep learning strategies in computer vision/image processing applications. The [3] demonstrates the efficiency of deep learning on simple image recognition while [14] addresses the 3D based object recognition problem. The use of conditional Deep Belief and Convolutional Networks for video sequence and human motion synthesis was reported in [15], [16]. However, none of the aforementioned approaches adopt an adaptive and fast deep learning strategy; therefore they cannot be applied in real-life application scenarios where dynamic and abrupt changes of the environment are encountered.

The proposed methodology assumes that few labeled data are first used to train multi-layered deep structures and then the “parameters” of the architecture are dynamically updated as new unlabelled data come to adjust the performance of the deep structures to the statistical characteristics of the new data. In this way, we are able to deal with the trade-off regarding adaptability versus stability. The proposed adaptable deep learning architectures are able to perform stable tracking even in complex environmental conditions, while retains the adaptable behavior of the tracker. The proposed framework incorporates highly non-linear models but simultaneously integrates optimization strategies able to update in real-time the non-linear model parameters according in a way to trust, as much as possible, the current changes of the environment, while providing a minimal degradation of the previous gained experience.

2 Problem Formulation

In the deep learning paradigm, a region instead of a single pixel value is used as input in object modeling. In particular, we consider the object detection problem as the estimation of a multi-valued non-linear function $\mathbf{h}(\cdot)$ that maps features of the region s

[we denote the respective feature map as **x**(s)] with a probability vector $\hat{\mathbf{d}}$ (an estimate of the actual probability vector **d**). Vector $\hat{\mathbf{d}}$ assigns a region *s* to one of the *M* available objects.

$$\hat{\mathbf{d}}(s) \approx \mathbf{h}_q(\mathbf{x}(s)) \tag{1}$$

where subscript q refers to the total parameters' space of the non-linear model $\mathbf{h}(\cdot)$. In the following, we assume, for simplicity, a two-class object detection problem. Then, the probability vector $\hat{\mathbf{d}}$ becomes scalar $d(s)$ and non-linear model as $h(\cdot)$. Extension to an *M* class object detection problem can be performed straightforwardedly.

The main difficulty in implementing Eq. (1) is that $h(\cdot)$ is actually unknown. Using concepts from functional analysis and assuming some simple restrictions regarding the continuity of $h(\cdot)$, we can model the $h(\cdot)$ as a finite sum of known functional components.

$$\hat{d}(s) \approx \phi(\mathbf{w}_N^T \cdot \mathbf{u}^{(N)}) \tag{2a}$$

$$\text{with } \mathbf{u}^{(l)} = \boldsymbol{\varphi}\left(\mathbf{W}_{l-1}^T \cdot \mathbf{u}^{(l-1)}\right),\ l = 1,2,\ldots,N \text{ and } \mathbf{u}^{(0)} = \mathbf{x}(s) \tag{2b}$$

In Eq. (2) $\boldsymbol{\varphi}(\cdot)$ is a known vector-valued functional component which should be bounded, continuous and monotonically increasing. In our case, we select function $\phi(\cdot)$ as the sigmoid function. Vector $\mathbf{w}_1$ are coefficients that weigh the non-linear transformations **u** of the input feature map **x**(*s*). Matrix $\mathbf{W}_0$ represents the weighted coefficients used for the initial transformation of **x**(*s*). *l* indicates the layer of the deep representation.

2.1 Limitations of Conventional Re-training

Although Eq. (2) is a robust mathematical framework for object modeling using deep architectures, the main difficulty results from the efficiency of the algorithm used to approximate the unknown coefficients $\mathbf{W}_l$ with *l*=1,…,*N*-1 and $\mathbf{w}_N$. In the following, for simplicity we denote all these coefficients as *q*. Under a supervised framework, a least squared steepest descent approach is usually applied to estimate the unknown coefficients *q*. However, in complex non-linear relationships and in case of a considerable number of input variables, there are multiple local minima in the error surface. Thus, it is quite possible the optimization to be trapped into local minima instead of global one. Another difficulty is that real-world applications are dynamic processes. The probabilistic characteristics of the data change through time while the collection effort for the labeled data is enormous and arduous. Furthermore, training a multi-layer structure on the use of only labeled data often leads to poor performance since adaptability is not permitted. Thus, supervised learning even though deep architectures are used, is not sufficient to address a stable object tracking for very long time spans, where usually abrupt and high visual changes take place. To address this problem, in this paper, we propose the mathematical framework of a fast and adaptive deep information fusing learning strategy and apply it to automatically update the behaviour of a tracker in a way to trust as much as possible, the current visual changes, while minimally degrading the already gained experience.

3 The Adaptive and Fast Deep Fusion Paradigm

Traditional classifiers use only labeled data (feature and label pairs) to train. Due to the nature of computer vision applications labeled data are difficult to obtain. This is due to the fact that it requires a specialist to label the data, which is not practically feasible especially under real-time video supervision processes. On the other hand, unlabeled data is abundant and can be easily collected. In this paper, we exploit this feature by letting the aforementioned deep learning structure to automatically self-adapt to the current conditions of the environment (dynamic changes of the visual scene conditions). Thus, the initially trained deep models can be enhanced, through the proposed adaptive strategy learning.

Let us now formulate the adaptive deep fusion framework. We assume that we have an incomplete (approximate) object model $h_{q_{in}}(\cdot)$ the parameters of which q_{in} have been estimated using for example a supervised training phase. Let us now assume that at a time *t*+*T* we process the image *I*(*t*+*T*) and extract a set of unlabeled features $\mathbf{x}_i^u(s)$ for regions $s \subseteq I(t+T)$. Then, our target is to refine the object function $h_{q_{in}}(\cdot)$ from the incomplete model coefficients q_{in} to a new more accurate (updated) model, i.e., $h_{q_{in}}(\cdot) \to h_{q_{ad}}(\cdot)$, exploiting the unlabeled data $\mathbf{x}_i^u(s)$.

In particular, initially, from the pool of unlabeled data *U* we estimate a very small set of samples of high confidence $C \subset U$ to belong to the objects of interest. Actually set C describes the current knowledge of the environment (see Section 4). Assuming a two-class object detection problem, set *C* can be exclusively divided into two sub-sets $C = C_1 \cup C_2$; where C_1 (C_2) refers to the foreground (background) object.

$$h_{q_{ad}}(\mathbf{x}_i^u) \approx B \quad \forall \mathbf{x}_i^u \in C \quad \text{where } B = \begin{cases} 0 & \forall \mathbf{x}_i^u \in C_1 \\ 1 & \forall \mathbf{x}_i^u \in C_2 \end{cases} \tag{3}$$

Eq. (3) means that after the adaptation the new coefficients q_{in} should be updated so that they trust as much as possible the most confident unlabeled data.

We also assume that a small perturbation of the coefficient space is sufficient to get an adequate modification of the model function. Therefore, we have that

$$q_{ad} = q_{in} + dq \text{ or equivalently} \tag{4}$$

$$\mathbf{W}_l^{ad} = \mathbf{W}_l^{in} + d\mathbf{W}_l \text{ for } l=1,..,N\text{-}1 \text{ and } \mathbf{w}_N^{ad} = \mathbf{w}_N^{in} + d\mathbf{w}_N \tag{5}$$

where dq , $d\mathbf{w}_N$, $d\mathbf{W}_l$ refers to a small perturbation of the coefficients space.

3.1 The Proposed Adaptive and Fast Retraining Strategy

Assuming a small perturbation for the coefficients, we can linearize Eq. (3) on the use of first order Taylor series expansion. This way, we can linearly express the unknown coefficients $d\mathbf{w}_N$ and $d\mathbf{W}_l$ as a function of the weights before adaptation $\mathbf{W}_l^{in}$ and $\mathbf{w}_N^{in}$ .Then, the following theorem can be proven.

Theorem 1: The constraint of (3) under the assumption of (5) is decomposed to a system of linear equations of the form $\mathbf{c} = \sum_l a_l^T \cdot vec(d\mathbf{W}_l)$, where vector **c** and vectors, $\mathbf{a}_l$ depends only on the previous known coefficients, $\mathbf{W}_l^{in}$ and $\mathbf{w}_N^{in}$; $vec(\cdot)$ is an operator that forms a vector from a matrix by stacking up all of the matrix elements.

Vector **c** expresses the differences between the multi-layer output after and before the adaptation. In other words, it provides an estimate of how much the classifier should be modified to trust the current visual properties (trust the current conditions). Vectors $\mathbf{a}_l$ are more complex relationships of the previous coefficients q_{in}. However, usually, a very small set of confident unlabeled data is selected to reduce error accumulation in tracking. Thus, the number of unknowns of the linear system of Theorem 1 is greater than the number of equations. To handle this problem, an additional constraint is introduced to restrict the solution of (3) on a feasible space. In our case, we assume that the norm $\|dq\|$ should undergone a minimal modification.

$$\min\|dq\| \tag{6}$$

The minimization of Eq. (6) subject to the constraint of Eq. (3) is in fact a convex minimization problem subject to linear constraints. Among several applicable techniques, the reduced gradient method has been selected due to its cost effectiveness.

4 Confident Data Selection

The purpose of this section is to automatically evaluate the unlabelled data so that the most confident ones are detected. This is an independent mechanism compared with the deep learning process; it exploits apart from the output of the object model (through the deep learning structure) additional criteria coming from moving coherency. It is worth noting that this process is not a classification framework, but an automatic way to train with unlabeled data. Let us form a graph $G = (V, E)$, the vertices of which corresponds to a set of selected unlabelled samples while the edge expresses a distance confident metric between two samples. In particular, let $e_{i,j}$ denote the graph edge between the node i and the node j. Then, graph edges should reflect the likelihood of the two samples to belong to the same object.

$$e_{i,j} = corr\left(\mathbf{x}(s_i),\mathbf{x}(s_j)\right)\cdot XNOR(\hat{d}(\mathbf{x}(s_i)),\hat{d}(\mathbf{x}(s_j)) \tag{7}$$

Eq. (7) expresses that the two samples present high likelihood to belong to the same objet class if their features present the same properties and the respective outputs of the initial (before updating) object model also present consistency. In case, we refer to foreground extraction applications, we can enrich (7) with additional constraints. For example, the likelihood that the two pixels belong to the foreground object is temporality constrained by the motion information. Thus, we have

$$Te_{i,j} = e_{i,j}\cdot AND(\vartheta(\mathbf{x}(s_i)),\vartheta(\mathbf{x}(s_j)) \tag{8}$$

To select the most confident unlabelled data, we partition the graph using spectral clustering algorithm.

5 Real-World Experiments

The experiments have been conducted using four public datasets; PETS2007 (a conference room), PETS2006 (a metro station) and two views of SCOVIS [19] depicting industrial workflows. Wide overlapped windows of 64x64 pixels have been chosen as regions s. Within this area, the MPEG-7 descriptors, such as Scalable Color, Dominant Color, and the Color Structure are used as features, which are fed into a 3-layer deep structure to perform tracking, each comprises of 10 neurons. In the rest, we focus on a foreground / background separation problem. The initial weights was estimated based on a training set of 320, 230, 570 samples for PETS2007, PETS2006 and SCOVIS dataset.

To estimate the most confident unlabelled data, we exploit the motion activity of the scene, estimated by Lucas-Kanade optical flow on selecting good features. For acceleration, we activate the adaptation strategy only in case where significant motion activity is encountered and we skip for the selection data of similar feature properties.

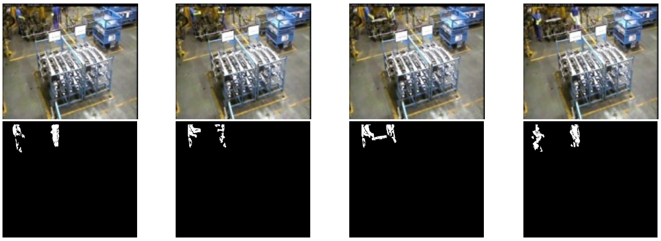


Fig. 2. Background changes effect. From left to Right: a,b) Two workers' detection of small size on complex background, c) tracking failure due to background change (workers move a car equipment), d) background content correction and accurate foreground detection by adaptation.

Initially, we present the results regarding SCOVIS dataset due to its complexity. To demonstrate tracker stability, we apply the proposed algorithm on more than

20,000 frames of SCOVIS other than the ones used in the training set. Within this long span of test sequence, several complex environmental changes occur which present challenges to maintain a stable object tracking. In particular, Fig. 2 presents a scene where a slight background change takes place; a car equipment is moving by the workers. As is observed, the tracker consistently detects the two workers. However, in Fig. 2c, the equipment is erroneously considered part of the tracking objects but as the algorithm runs and more unlabelled data are exploited, the proposed deep structure unlearns the error and sets the equipment as background (Fig. 2d).

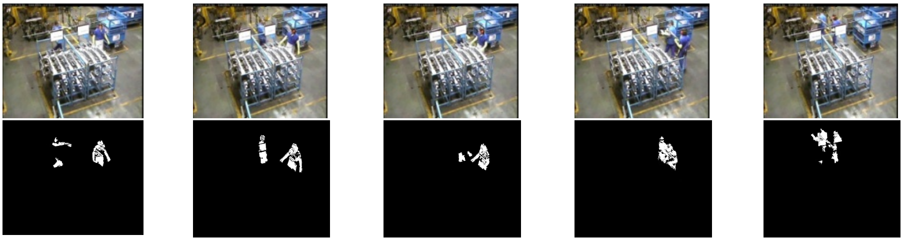


Fig. 3. Occlusions effect. From left to Right: a) The left worker is partially occluded and the tracker stably detects this, b) the left worker re-appears, c) again the left worker is occluded, d) the two workers overlaps each other, e) both workers are partially occluded.

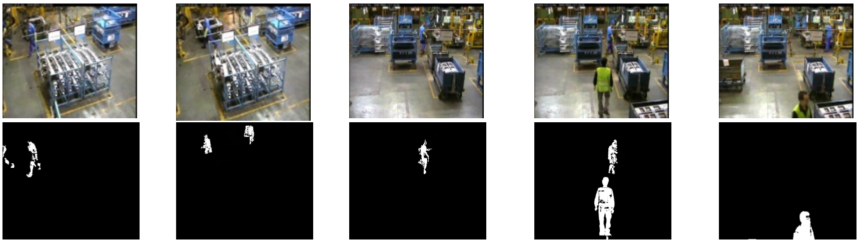


Fig. 4. Tracking effect over long time spans. a) One of the two workers disappears from the scene, while the background has changed from Fig. 3. b) Welding fire in the background; tracking efficiency under illumination changes. c) Active camera effect; the background has significantly changed; stable tracking after several adaptation cycles. d,e) new foreground enters the scene, stable tracking after several iterations, through the second worker is missed in Fig 4e.

Fig. 3 shows five frames of another part of the sequence to demonstrate the effect on occlusions. In Fig. 3a, the left worker is partially extracted since he is occluded by the rack. Then, he is tracked again (Fig. 3b) while he is again occluded in Fig. 3c. Similar performance is observed for the other frames. In Fig. 4a, we observe the stability of the algorithm as the worker leaves the scene. Fig. 4b show the stability of the algorithm for high illumination changes (fire welding). Finally, the last two frame presents tracking performance for long time spans where background has significantly change and new foreground object enter the scene. After several adaptation cycles stable tracking is encountered.

We then objectively evaluate and compare our scheme with other methods. The evaluation was performed using four segments of the SCOVIS sequence of different visual properties, which are merged together to form one sequence. Fig. 5 shows the confident error as the ratio of the XOR of the tracked and ground truth mask over ground truth. In Fig. 5, we have compared the results with the on-line SVM [10], the Semi Boost [8], the Students-t [18] and the Gaussian [17] tracking. For fair comparison, we have modified the algorithms to use the same confident unlabelled data. We observe that our algorithm better handles the problem of adaptability vs. stability in terms of providing well enough generative models which can simultaneously specialize well to visual changes. Table 1 shows the average confidence over the four examined video sequences of our method and the four compared ones.

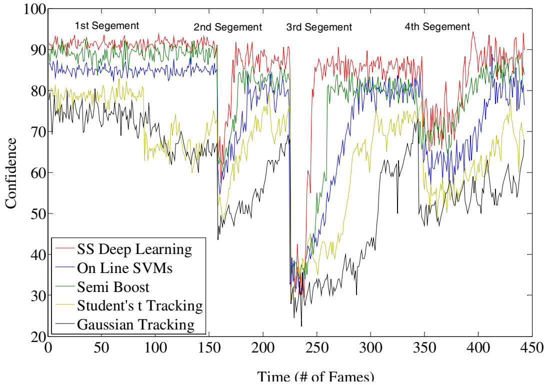


Fig. 5. Comparison of the proposed algorithm with other adaptable tracking methodologies over a merged of difference scene segments of SCOVIS dataset. We observe that the proposed algorithm is more stable in terms of adaptability and tracking accuracy.

Table 1. Average tracking confidence over different datasets and methods

Sequences	**SSL Deep Learning**	**On Line SVM**	**Semi Boost**	**Student's t Distribution**	**Gaussian Tracking**
SCOVIS Cam 32	85.22%	76.28%	80.32%	65.99%	58.02%
SCOVIS Cam 34	84.64%	74.66%	79.73%	65.23%	57.57%
PETS2007	88.33%	82.22%	84.44%	79.55%	74.54%
PETS2006	90.12%	87.33%	89.12%	83.22%	80.18%

6 Conclusions

In this paper, we introduced a semi-supervised deep learning algorithm for stable long time object tracking in real-time. We exploited perturbation theory with optimization strategies to efficiently self-adapt non-linear deep structures in a way to trust as much as possible the current visual properties, while simultaneously providing a minimal degradation of the already gained experience. We have tested the proposed semi-supervised deep learning under quite complex video footages, where several occlusions, illumination changes, background / foreground content modification are encountered.

Acknowledgements: The research leading to these results has been supported by European Union funds and national funds from Greece and Cyprus under the project ”POSEIDON: Development of an Intelligent System for Coast Monitoring using Camera Arrays and Sensor Networks” in the context of the inter-regional programme INTERREG (Greece-Cyprus cooperation) - contract agreement K1 3 10–17/6/2011.

References

[1] Stalder, S., Grabner, H., van Gool, L.: Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition. In: Proc. of IEEE ICCV, pp. 1409–1416 (2009)

[2] Henriques, J.F., Caseiro, R., Batista, J.: Globally optimal solution to multi-object tracking with merged measurements. In: Proc. of IEEE ICCV, pp. 2470–2477 (2011)

[3] Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation 18, 1527–1554 (2006)

[4] Lepetit, V., Lagger, P., Fua, P.: Randomized trees for real-time keypoint recognition. In: Proc. IEEE CVPR, vol. 2, pp. 775–781 (2005)

[5] Collins, R.T., Liu, Y.: On-line selection of discriminative tracking features. In: Proc. of IEEE ICCV, vol. 1, pp. 346–352 (2003)

[6] Doulamis, A., Ntalianis, K., Doulamis, N., Kollias, S.: An Efficient Fully-Unsupervised Video Object Segmentation Scheme Using an Adaptive Neural Network Classifier Architecture. IEEE Trans. on NNs 14(3), 616–630 (2003)

[7] Matthews, L., Ishikawa, T., Baker, S.: The template update problem. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(6), 810–815 (2004)

[8] Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)

[9] Tang, F., Brennan, S., Zhao, Q., Tao, H.: Co-tracking using semi-supervised support vector machines. In: Proc. of IEEE ICCV, pp. 1–8 (2007)

[10] Yu, Q., Dinh, T.B., Medioni, G.: Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 678–691. Springer, Heidelberg (2008)

[11] Cehovin, L., Kristan, M., Leonardis, A.: An adaptive coupled-layer visual model for robust visual tracking. In: Proc. of IEEE ICCV, pp. 1363–1370 (2011)

[12] Xing, J., Ai, H., Lao, S.: Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1200–1207 (2009)

[13] Zhang, L., Yuan, L., Nevatia, R.: Global data association for multi-object tracking using network flows. In: Proc. of IEEE CVPR, pp. 1–8 (2008)

[14] Nair, V., Hinton, G.: 3-D object recognition with deep belief nets. In: Proc. NIPS (2009)

[15] Taylor, G., Hinton, G.E., Roweis, S.: Modeling human motion using binary latent variables. In: Proc. NIPS (2007)

[16] Schulz, H., Behnke, S.: Object-Class Segmentation using Deep Convolutional Neural Networks. In: Proc. of DAGM Workshop, Frankfurt (August 2011)

[17] Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Proc. of IEEE CVPR, Fort Colins, CO (June 1999)

[18] Moghaddam, Z., Piccardi, M.: Robust density modelling using the student’s t-distribution for human action recognition. In: Proc. of IEEE ICIP, pp. 3261–3264 (2011)

[19] Voulodimos, A., Kosmopoulos, D.I., Vasileiou, G., Sardis, E., Doulamis, A.D., Anagnostopoulos, V., Lalos, C., Varvarigou, T.A.: A dataset for workflow recognition in industrial scenes. In: Proc. of IEEE ICIP, pp. 3249–3252 (2011)