Abstract: Technologies of Vision (TeV) conducts research in computer vision, image analysis and visual pattern recognition. Main activities concern the development of computer vision methods for semantic image labelling and dynamic scene understanding with application in video analytics and augmented reality. We also investigate on machine learning techniques for image segmentation, classification and retrieval. With our activity, we aim at advancing the state-of-the-art in computer vision research and technology with particular attention to applicability in practice and integration. We also put effort in the realization of demonstrators and experimentation, to seek for further valorization opportunities of research results. TEV is a research unit of the Center for Information Technology of Fondazione Bruno Kessler.

Key words: Dynamic scene understanding, Vision for AR, Semantic image labelling, Color analysis

1. INTRODUCTION

TeV main activities focus on: (i) development and experimentation of computer vision methods for monitoring people in indoor spaces including real time people tracking, head pose estimation, person re-identification, action recognition and human behavior analysis; (ii) development and experimentation of computer vision methods for augmented reality including image/mobile camera registration and scene text/visual pattern recognition;
(iii) design of novel machine learning methods for semantic image labeling featuring joint data representation and discriminative learning from sparsely annotated data; (iv) development of color correction methods based on color constancy principle for pre-processing color vision algorithms including application to object recognition and tracking.

We envision the convergence of the first two activities into the Smart Spaces of the Future, where real-time behavioral attention and movement sensing by advanced machine vision enables intelligent augmented reality services such as contextualized information access and guidance. Methodological investigations on machine learning with unstructured input will lay down the basis for new endeavors in video and media analytics that require the synergic integration across multiple modalities with high-level interpretation and reasoning mechanisms.

2. DYNAMIC SCENE UNDERSTANDING

In this section we present the major recent advancements on real time people tracking [5-6] and head pose estimation [1-2] from multiple surveillance cameras, its use for human behavior analysis [3], statistical methods for action recognition [4] and anomaly detection [7]. Moreover, in the context of an industrial project in the field of video surveillance, novel algorithms have been implemented to deal with the person re-identification problem, achieving results that outperform the current state-of-the-art. The abstract of each mentioned work is reported in the following.


Abstract. Head pose classification from surveillance images acquired with distant, large field-of-view cameras is difficult as faces are captured at low-resolution and have a blurred appearance. Domain adaptation approaches are useful for transferring knowledge from the training (source) to the test (target) data when they have different attributes, minimizing target data labeling efforts in the process. This paper examines the use of transfer learning for efficient multi-view head pose classification with minimal target training data under three challenging situations: (i) where the range of head poses in the source and target images is different, (ii) where source images capture a stationary person while target images capture a moving person whose facial appearance varies under motion due to changing perspective, scale and (iii) a combination of (i) and (ii). On the whole, the presented methods represent novel transfer learning solutions employed in the context of multi-view head pose classification. We demonstrate that the proposed solutions considerably outperform the state-of-the-art through extensive experimental
validation. Finally, the DPOSE dataset compiled for benchmarking head pose classification performance with moving persons, and to aid behavioral understanding applications is presented in this work.


Abstract. We propose a novel Multi-Task Learning framework (FEGA-MTL) for classifying the head pose of a person who moves freely in an environment monitored by multiple, large field-of-view surveillance cameras. As the target (person) moves, distortions in facial appearance owing to camera perspective and scale severely impede performance of traditional head pose classification methods. FEGA-MTL operates on a dense uniform spatial grid and learns appearance relationships across partitions as well as partition-specific appearance variations for a given head pose to build region-specific classifiers. Guided by two graphs which a-priori model appearance similarity among (i) grid partitions based on camera geometry and (ii) head pose classes, the learner efficiently clusters appearance-wise related grid partitions to derive the optimal partitioning. For pose classification, upon determining the target’s position using a person tracker, the appropriate region-specific classifier is invoked. Experiments confirm that FEGA-MTL achieves state-of-the-art classification with few training data.


Abstract. Correlates between social attention and personality traits have been widely acknowledged in social psychology studies. Head pose has commonly been employed as a proxy for determining the social attention direction in small group interactions. However, the impact of head pose estimation errors on personality estimates has not been studied to our knowledge. In this work, we consider the unstructured and dynamic cocktail party scenario where the scene is captured by multiple, large field-of-view cameras. Head pose estimation is a challenging task under these conditions owing to the uninhibited motion of persons (due to which facial appearance varies owing to perspective and scale changes), and the low resolution of captured faces. Based on proxemic and social attention features computed from position and head pose annotations, we first demonstrate that social attention features are excellent predictors of the Extraversion and Neuroticism personality traits. We then repeat classification experiments with behavioral features computed from automated estimates—obtained experimental results show that while prediction performance for both traits is affected by head pose estimation errors, the impact is more adverse for Extraversion.


Abstract. The widespread adoption of low-cost depth cameras has opened new opportunities to improve traditional action recognition systems. In this paper, we focus on the specific problem of action recognition under viewpoint changes and propose a novel approach for view-invariant action recognition operating jointly on visual data of color and depth camera channels. Our method is based on the unique combination of robust Self-Similarity Matrix (SSM) descriptors and multi-task learning. Indeed, multi-view action recognition is inherently a multi-task learning problem: images from a camera view can be
modeled as visual data associated to the same task and it is reasonable to assume that the data of different tasks (camera views) are related to each other. In this work we propose a novel algorithm extending Multi-Task Linear Discriminant Analysis (MT-LDA) to enhance its flexibility by learning the dependencies between different views. Extensive experimental results on the publicly available ACT42 dataset demonstrate the effectiveness of the proposed method.


Abstract. People tracking under non-uniform illumination is challenging, as observed appearance may change as they move around in the environment. Appearance model adaptation is inconvenient over the long run as it is subject to drift, while filtering illumination information in the data through built-in invariance is sub-optimal in terms of discriminative capability. In this work, we are interested in modeling the spatial and temporal dimensions of appearance variation induced by non-uniform illumination, and to learn and adapt related parameters over time by using walking people as illumination probes. We propose a hybrid graphical model and a new message passing scheme that sequentially updates parameters of the model, so that scene illumination can be learnt online and used for robust tracking in dynamic environment.


Abstract. The employment of visual sensor networks in surveillance systems has brought in as many challenges as advantages. While the integration of multiple cameras into a network has the potential advantage of fusing complementary observations from sensors and enlarging visual coverage, it also increases the complexity of tracking tasks and poses challenges to system scalability. A key approach to tackling these challenges is the mapping of the demanding global task onto a distributed sensing and processing infrastructure. In this paper, we present an efficient and scalable multi-camera multi-people tracking system with a three-layer architecture, in which we formulate the overall task (i.e. tracking all people using all available cameras) as a vision based state estimation problem and aim to maximize utility and sharing of available sensing and processing resources. By exploiting the geometric relations between sensing geometry and people’s positions, our method is able to dynamically and adaptively partition the overall task into a number of nearly independent subtasks, each of which tracks a subset of people with a subset of cameras. The method hereby reduces task complexity dramatically and helps to boost parallelization and maximize the real time throughput and available resources of the system while accounting for intrinsic uncertainty induced, e.g., by visual clutter, occlusion, and illumination changes. We demonstrate the effectiveness of our method by testing it with a challenging video sequence.


Abstract. In the last decades, many efforts have been devoted to develop methods for automatic scene understanding in the context of video surveillance applications. This paper presents a novel non-object centric approach for complex scene analysis. Similarly to previous methods, we use low-level cues to individuate atomic activities and create clip histograms. Differently from recent works, the task of discovering high-level activity patterns is formulated as a convex prototype learning problem. This
problem results in a simple linear program that can be solved efficiently with standard solvers. The main advantage of our approach is that, using as the objective function the Earth Mover's Distance (EMD), the similarity among elementary activities is taken into account in the learning phase. To improve scalability we also consider some variants of EMD adopting L1 as ground distance for 1D and 2D, linear and circular histograms. In these cases, only the similarity between neighboring atomic activities, corresponding to adjacent histogram bins, is taken into account. Therefore, we also propose an automatic strategy for sorting atomic activities. Experimental results on publicly available datasets show that our method compares favorably with state-of-the-art approaches, often outperforming them.

3. VISION FOR AUGMENTED REALITY

R&D activities on computer vision for augmented reality fall within the context of VENTURI FP7 project (http://venturi.fbk.eu). Our research focus is on visual pattern recognition and vision-assisted registration and orientation tracking of on-board camera for indoor/outdoor augmented reality on smart-phone and tablet. The main achievements of the 2nd year activities have been integrated into the demonstrator for the “Indoor Navigation and Assistant for the Visually Impaired” use-case. We contribute to the demonstrator development at various levels: module integration, user communication, and implementation of computer vision algorithms. In particular, we developed several low level modules, called target detectors, specialized to the detection and tracking of specific visual targets, and some modules for visual navigation that exploit the target detectors to recognize landmarks embedded in the environment.

3.1 Visual target detector

A visual target detector is a module able to detect and to localize, within the scene acquired by the smartphone camera, a specific target. The target is composed by one or more parts and its structure is coded in a specific description file (xml format). The description file stores information about the number of parts, their shape/size in the real world (represented by polylines) and the spatial relationships among them. The different parts are assumed coplanar. For each part, a specialized routine can be specified in the xml file that is applied to the input frame to obtain a list of regions that are likely to depict the part at hand.

For the second year use case we have developed eight detectors, six of them are composed by a single part (WetFloor, FBK_panel, AR_Logo, LiftSign, ToyShop, LegoLogo) and two are composed by several parts (InsideLiftButtons, OutsideLiftButtons), see Figure 1.
An additional detector (Fingertip) has been implemented to detect the user fingertip while touching the facing lift buttons pad. Thanks to the homography computed by the buttons detector it is possible to estimate the relative position, in real world coordinates, of the fingertips with respect to the button to be pressed. The introduction of a multi-part target allows detectors to be robust to partial occlusions, a useful feature especially in the case of detecting lift buttons covered by the user hand.

### 3.2 Visual navigators

A visual navigator coordinates one or more detectors and interacts with the (visually impaired) user in order to instruct him/her to safely reach a predefined destination. The interaction is managed by means of spoken messages (from system to user) and a smart-watch (bidirectional). The visual navigator receives the list of detectors to be activated, in sequence or in parallel, in order to reach a set of local destinations. It launches the target detectors in a proper order and manages in real time their outputs. According to them, it notifies the user as soon as the target is detected in the scene, it communicates the direction to follow in order to reach it, and the action to perform to pass from one target to the next. Moreover, the navigator warns the user in the case that specific situations are detected (i.e. wet floor sign).
Four visual navigators have been developed for the second year demonstration of Venturi.

The purpose of the first one is to guide the user from the FBK entrance, which simulates the entrance of a shopping mall, towards the lift and to push the lift button. Along the route, it checks for the presence of a wet floor sign and in that case warns the user. The involved detectors are: *WetFloor, FBK_panel, AR_logo, LiftSign, OutsideLiftButtons* and *Fingertip*. At the beginning, *FBK_panel* and *WetFloor* detectors are activated in parallel and the user is invited to scan the scene until the FBK stand banner is captured by the smartphone camera. Once detected the user is guided towards it by means of spoken messages with the purpose to maintain the target in the middle part of the image. When the estimated distance between the user and the target falls below two meters the visual navigators launches the *AR_logo* detector. In a similar way, the system looks for the target and guides the user towards it. Again, when the distance is below 0.8 meters a new detector (*LiftSign*) is activated and the user instructed how to aim the camera in order to frame the lift sign. When the estimated distance of the target falls below 2 meters, the user is in the area close to the lift. It is invited to turn right to search for the buttons pad and the correspondent detector is activated (*OutsideLiftButtons*). When the user is close enough (less than 0.5 meters) is invited to touch the buttons pad with a finger and the *Fingertip* detector is enabled.

![Figure 2: THE FIRST VISUAL NAVIGATOR AT WORK](image)

The system estimates the position of the fingertip with respect to the "arrow up" button and instructs the user how to move the finger until it goes
over the correct button. Eventually the visual navigator invites the user to press the button and to wait for the lift. Figure 2 shows some snapshots taken during the execution of the first visual navigator.

When the user enters the lift and taps the smart-watch, the second visual navigator is activated. Its purpose is to guide the user finger in order to reach and push the "Floor 1" button, where the toys shop is located. It uses only the \textit{InsideLiftButtons} detector with the Fingertip detector enabled when appropriate.

The third visual navigator is activated when the user is in the proximity of the toyshop. It launches the ToyShop detector and guides the user toward the shop entrance.

Finally, the fourth visual navigator is activated when the user confirms to be inside the shop by tapping the smart-watch. The LegoLogo detector is launched and the navigator guides the user towards the logo boxes shelf. Figure 3 shows some snapshots taken during the execution of the second, third and fourth visual navigator.

![Figure 3: PICTURES TAKEN DURING THE EXECUTION OF THE SECOND, THIRD AND FOURTH VISUAL NAVIGATOR, RESPECTIVELY](image)

4. **SEMANTIC IMAGE LABELLING**

We achieve novel results in generative-discriminative learning with Neural Decision Forests [8], a novel approach to jointly tackle data representation and discriminative learning within randomized decision trees. The key findings in our experiments on three different semantic image labelling datasets, are consistently improved results and significantly compressed trees compared to conventional classification trees.


**Abstract.** In this work we present Neural Decision Forests, a novel approach to jointly tackle data representation- and discriminative learning within randomized decision trees. Recent advances of deep learning architectures demonstrate the power of embedding representation learning within the classifier – an idea that is intuitively supported by the hierarchical nature of the decision forest model where the
Overview of Research Activities at Technologies of Vision Research
Unit of Fondazione Bruno Kessler

input space is typically left unchanged during training and testing. We bridge this gap by introducing randomized Multi-Layer Perceptrons (rMLP) as new split nodes which are capable of learning non-linear, data-specific representations and taking advantage of them by finding optimal predictions for the emerging child nodes. To prevent overfitting, we i) randomly select the image data fed to the input layer, ii) automatically adapt the rMLP topology to meet the complexity of the data arriving at the node and iii) introduce an $l_1$-norm based regularization that additionally sparsifies the network. The key findings in our experiments on three different semantic image labelling datasets are consistently improved results and significantly compressed trees compared to conventional classification trees.

5. COLOR ANALYSIS

Further work addresses the problem of estimating a color correction, between images observing the same scene but acquired under different illuminants and/or by different devices, that can be applied as a pre-processing step to color vision processing such as image matching and object recognition [9-10] and tracking [5].


Abstract. This work presents a novel color correction algorithm between images (or image regions) depicting the same environment. Color transfer is achieved by a linear full $3 \times 3$ mapping between the color responses. In particular, this linear map equalizes the colors of images of the same scene acquired under different illuminants and/or by different devices. Since the complexity of the proposed method is directly proportional to the number of picture pixels, color correction is performed in linear time within the pictures size.


Abstract. This work describes a novel color pixel topology that converts the three chromatic components from the standard RGB space into the normalized $r\cdot g$ chromaticity space. This conversion is implemented with high-dynamic range and with no dc power consumption, and the auto-exposure capability of the sensor ensures to capture a high quality chromatic signal, even in presence of very bright illuminants or in the darkness. The pixel is intended to become the basic building block of a CMOS color vision sensor, targeted to ultra-low power applications for mobile devices, such as human machine interfaces, gesture recognition, face detection. The experiments show that significant improvements of the proposed pixel with respect to standard cameras in terms of energy saving and accuracy on data acquisition. An application to skin color-based description is presented.