Clothes Change Detection Using the Kinect Sensor

Dimitrios Sgouropoulos, Theodoros Giannakopoulos, Sergios Petridis,
Stavros Perantonis and Antonis Korakis

Computational Intelligence Laboratory (CIL), Institute of Informatics and Telecommunications, National Center for Scientific Research DEMOKRITOS, Patriarchou Grigoriou & Neapoleos, Ag. Paraskevi 15310, Athens, Greece

Keywords: Kinect, Clothes Change Detection, Activities of Daily Living.

Abstract: This paper describes a methodology for detecting when a human has changed clothes. Changing clothes is a basic activity of daily living which makes the methodology valuable for tracking the functional status of elderly people, in the context of a non-contract unobtrusive monitoring system. Our approach uses Kinect and the OpenNI SDK, along with a workflow of basic image analysis steps. Evaluation has been conducted on a set of real recordings under various illumination conditions, which is publicly available along with the source code of the proposed system at http://users.iit.demokritos.gr/ tyianak/ClothesCode.html.

1 INTRODUCTION

The elderly population is constantly growing during the last decades and is expected to grow dramatically over the next few years, especially in Europe. This increase has made elderly care a rapidly growing task and in particular it has led to major research effort on implementing automatic assistive services for the elderly, in order to facilitate independent living. In this work, we employ image analysis techniques applied on data recorded from the Kinect sensor (Kin, 2011)(Zhang, 2012), in order to detect that a human has changed clothes between two successive recording sessions. The purpose of such a service is to measure the functional status of a person in the context of in-home unobtrusive health monitoring. The ability to change clothes is an important self-care tasks taken into consideration by health professionals when monitoring a patient, especially for the case of people with disabilities or the elderly. Such functional activities are referred to as *Activities of daily living (ALDs)* (Self-maintenance, 1969) and include self-care tasks such as: bathing, personal hygiene, toilet hygiene and eating (Collin et al., 1988)(Collin and Wade, 1988)

Automatically recognizing ADLs has gained research interest during the last years. This is usually achieved through sensors such as accelerometers, Radio Frequency Identification (RFIDs), microphones and cameras (Fleury et al., 2010)(Stikic et al., 2008). In (Fleury et al., 2010) a multi-class Support Vector Machine (SVM) has been employed in order to recognize among 7 ADLs, based on several sensors: infra-red presence sensor, wearable kinematic sensors, microphones and others. For the particular case of the dressing/undressing activity, the overall classification accuracy was found equal to 75%, while the maximum confusion was observed for the “resting” and “sleeping” activities. Instead of recognizing the (un)dressing activity among other classes of events, in this work we focus on simply answering the *binary* question: “has the person changed clothes between two successive recordings?”. The task then simplifies to (a) detect the clothes worn by the person (b) model the clothes and (c) measure the similarity of clothes detected between two recordings.

Automatically recognizing apparel using visual information can have a wide range of potential applications: surveillance, e-commerce, household automated services, etc. Depending on the field of application and the approach of information acquisition the respective methods has been either applied on single color images (e.g., (Chen et al., 2012)(Liu et al., 2012b)) or combination of color and depth images (e.g., (Maitin-Shepard et al., 2010)). In the context of shopping recommendation and customer profiling, some papers have proposed adopting visual analysis methods to describe clothing appearance with semantic attributes. (Chen et al., 2012) proposes using SIFT descriptors and SVMs to predict 26 predefined attributes concerning clothing patterns, colors, gender as well as general clothing categories (e.g. shirts). (Liu et al., 2012b) describes a method for

Sgouropoulos D., Giannakopoulos T., Petridis S., Perantonis S. and Korakis A..
Clothes Change Detection Using the Kinect Sensor.
DOI: 10.5220/0005001200850089
In *Proceedings of the 11th International Conference on Signal Processing and Multimedia Applications* (SIGMAP-2014), pages 85-89
ISBN: 978-989-758-046-8

clothing retrieval using a daily human photo captured in a general environment (e.g. street). In (Liu et al., 2012a), an automatic occasion-oriented (e.g. wedding) clothing recommendation system is presented. Towards this end, Histograms of Oriented Gradient (HOG) and color histograms have been adopted as features, while SVMs have been used as classifiers. In a similar context, (Bossard et al., 2013) introduces a pipeline for recognizing and classifying people's clothing in natural scenes. Among others, HOGs and color histograms are used as features, while classification is achieved via Random Forests. (Kalantidis et al., 2013) presents a visual analysis method that suggests clothing results given a single image.

Visual-based clothing classification has also been used in household service robotic applications, i.e. in automated laundry. In (Willimon et al., 2011), a robotic system which identifies and extracts items sequentially from a pile using only visual sensors is described. Classification of each clothing item is conducted based on a six-class hierarchy (pants, shorts, short-sleeve shirt, long-sleeve shirt, socks, or underwear). Depth information is also used based on a stereo pair of cameras. In (Maitin-Shepard et al., 2010) an application of robotic towel folding is presented, where image (both color and depth) analysis is adopted to detect corners that can be used for grasping the towel. (Ramisa et al., 2012) also presents a grasping point detection method using color and depth information from a Kinect device. This work focuses on identifying the grasping points in one single step, even when clothes are highly wrinkled, therefore avoiding multiple re-graspings. (Willimon et al., 2013) focuses on defining mid-level features in order to boost the clothing classification performance. Again, the task here is to classify clothes from a pile of laundry (three categories have been used: shirts, socks and dresses).

2 METHODOLOGY

2.1 Clothes Detection

The Kinect Sensor provides RGB, depth data and skeletal tracking information, i.e 3D coordinates of tracked body skeletons ((Xia et al., 2011)(Shotton et al., 2013)). In particular, we have employed the OpenNI SDK (http://www.openni.org/) in order to identify the positions of these key joints on the human body (hands, elbows, head, etc), along with other human position information such as orientation estimates and distance from the sensor. The first step is to specify the bounds of the areas of interest to narrow down the particular task. The Kinect middle-ware produces two matrices that are used for this purpose: (a) pixel matrix: this matrix contains the color information of each pixel in the RGB color space (b) user matrix: this matrix indicates if the respective pixels belong to a user (human) or not.

Using these matrices it is possible to maintain only the important information, that is, the user-related color values. Following that basic notation, the areas of interest for the particular task have been set. In particular, two primary areas have been defined, namely, the torso area (used to model the upper clothes evaluation) and the lower body (used to model the lower clothes). The determination of these areas of interest was based on the user data provided by Kinect and information stemming from particular joint coordinates of the skeleton, also provided by the Kinect sensor. The upper body area is based on a rectangular area, with dimensions that are primarily defined by the shoulders and torso joints and then enhanced based on the user's body width and height. Following the same method we form the user's lower body area by using his hip and knee joints, from the skeleton estimate, and performing similar improvements.

2.2 Clothes Color Representation

For each area of interest (torso and lower body), 60 feature values related to the color of the respective clothing are extracted. In particular, 30 features stem from the three histograms from the color information (RGB), since 10 bins per color channels are used in the histogram calibration. Similarly, 30 features stem from the histograms of the edges of each color coordinate. Towards, this end, the Sobel image operator is applied on each color coordinate. At each frame where a human is detected, a feature vector of the area of interest is calculated as described above, along with a respective confidence measure related to that detection. This process forms a feature matrix $X : M \times D$, whose rows correspond to the respective feature vectors. The confidence measure is extracted according to the following weighted heuristic:

$$H(O,R,D) = w_1 \cdot cos^2(O) + w_2 \cdot R + w_3 \cdot \frac{1}{2\sqrt{\pi\sigma}} e^{-\frac{D^2}{2\sigma^2}} \tag{1}$$

The first factor is based on the user's orientation in the room (in degrees): frontal orientation either looking to the Kinect ($O = 0$) or at the other side ($O = 180$) gives the highest confidence while profile orientations (e.g. $O = 90$) gives the lowest. The second factor depends on R which is the ratio of the current to the previous user pixel count (i.e. number of pixels that belong to the human, as estimated by the middleware). The last factor takes into account the user's distance from the sensor, where the σ is set to reflect the expected. The respective weights of each factor w_i are determined by the efficiency of the heuristic in our various experiments and our only restriction is that $w_1 + w_2 + w_3 = 1$. In our experiments we use $w_1 = w_2 = 0.4$, $w_3 = 0.2$.

Finally, each recording session is represented as a single feature vector which is computed as a weighted average of individual feature vectors $F_n = \frac{\sum_{i=1}^{M} X_{i,n} \cdot C_i}{\sum_{i=1}^{M} C_i}$, where D is the number of feature dimensions (60), M is the number of samples (feature vectors) of the recording session, $X_{i,n}$ is the n-th feature of the i-th sample, $n = 1, \ldots, D$, and C_i is the confidence value of the i-th sample of the recording.

2.3 Color Constancy

The proposed system should function under a real home-environment, therefore there is a need for robustness to varying illumination conditions. Towards this end, we include in our methodology a color constancy method, aiming towards color features retaining constant statistics under different illumination conditions (Funt et al., 1996). In particular, we have experimented with the following static methods for color constancy: (a) the Grey-World algorithm which assumes that the average color in a scene is achromatic and therefore normalizes each color based to the respective gray (average) values (b) the White-Patch method which normalizes each color coordinate by the respective maximum channel value achieving maximization towards a hypothetical white reference area and (c) a simple modification of the White-Patch which uses the average value of a range of high-valued pixels (instead of using the global maximum) in order to increase robustness to noise.

3 EXPERIMENTS

3.1 Data Used

In order to evaluate the cloth change detection ability of the proposed approach, a dataset of real recordings has been compiled and manually annotated. In total, four humans have participated in the recordings under two different lighting conditions, namely natural and artificial lighting. For each case, a different number of upper and lower apparel has been used. Each recorded session is stored on a separate `oni` file using the OpenNI library. The name of these files indicate the IDs of the corresponding apparel.

3.2 Evaluation Method

In this Section we describe the adopted methodology for the evaluation of the discrimination ability of the adopted color representation. We particularly describe the process for the upper clothes as it is exactly the same for the lower clothes case. Given:

- a set of upper clothes feature vectors $\mathbf{FU_i}$, $i = 1, \ldots, N$, where N is the total number of video sessions of 60 elements each.
- a vector of upper labels LU_i, $i = 1, \ldots, N$ where each different value represents a distinct piece of clothing. This is used as ground truth in the evaluation process.

We start by creating the confusion matrix CM and initializing it with zeros. Then for each possible pair of $\mathbf{FU_i}$ and $\mathbf{FU_j}$ where $i = 1, \ldots, N$ and $j = 1, \ldots, N, j \neq i$ we compute their Euclidean distance $DU_{i,j}$ and compare it to a user-defined threshold T. If $DU_{i,j}$ is greater than T then that pair of clothes is perceived to be different, otherwise the same. So we now have four separate cases:

- $CM_{1,1}$: number of times that two feature vectors have the same estimated label ($DU_{i,j} \leq T$) and the same ground truth label ($LU_i = LU_j$) - true negative
- $CM_{1,2}$: number of times that two feature vectors have different estimated labels ($DU_{i,j} > T$) but the same ground truth label ($LU_i = LU_j$) - false positive
- $CM_{2,1}$: number of times that two feature vectors have the same estimated label ($DU_{i,j} \leq T$) but different ground truth labels ($LU_i \neq LU_j$)- false negative
- $CM_{2,2}$: number of times that two feature vectors have different estimated labels ($DU_{i,j} > T$) and different truth labels ($LU_i \neq LU_j$) - true positive

After computing the overall confusion matrix, as described above, it is normalized so that the two events are considered equiprobable and finally the performance measures Precision, Recall and F1 measure are calculated: $Pr = \frac{CM_{2,2}}{\sum_{i=1}^{2} CM_{i,2}}$, $Re = \frac{CM_{2,2}}{\sum_{i=1}^{2} CM_{2,i}}$ and $F1 = \frac{2 \cdot Pr \cdot Re}{Pr + Re}$. The exact same evaluation process is repeated for the lower clothing.

3.3 Evaluation Results

As described above, the recordings have been conducted under two different general illumination cate-

Table 1: $F1$ evaluation results (%) for different lighting conditions and all feature calibration methods.

	Artificial	Natural	Mixed
No color constancy	77	82	72
Gray world	78	83	71
White Patch	84	82	77
Modified White Patch	85	85	80

gories (natural and artificial lighting). The evaluation has been based on these two categories, as long as their "mixed" condition: the latter is the general (and harder) case of detecting changes under all possible illumination conditions. The results of this process are shown in Table 1. Due to space limitations, we do not present the performance results for the upper and lower clothings but only their averages. However, we would like to report that, in average, the problem of detecting changes on the lower clothes is at least 10% harder, in terms of $F1$ measure. This is probably due to the fact that the lower body part is usually not entirely visible, in the context of a real home environment, since there are usually pieces of furniture and other objects intervening between the sensor and the human.

4 CONCLUSIONS

We have presented a Kinect-based approach to detecting changes in users' clothes in a smart home environment in the context of measuring the functional status of the elderly. The whole system has been implemented in the Processing programming language, using the OpenNI SDK and achieves real-time detection. In order to evaluate the proposed approach, a dataset of recordings under various illumination conditions has been compiled, which is also publicly available. Experimental results have indicated that the overall change detection method achieves up to 80% performance for mixed lighting conditions and 85 for single conditions, that is 8% compared to the performance when the initial feature representation is adopted. In addition, the adopted color constancy approach abridges the gap at the performance between different illumination conditions. In the context of the carried out ongoing work we focus on the following directions: (a) implementation of more advanced image features (e.g. HOGs) (b) evaluation of more sophisticated color constancy techniques and (c) extension of the benchmark with more users and clothes combinations.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 288532. For more details, please see http://www.usefil.eu.

REFERENCES

(2011). Microsoft kinect sensor. Online available: http://www.microsoft.com/en-us/kinectforwindows/. Accessed April 1, 2013.

Bossard, L., Dantone, M., Leistner, C., Wengert, C., Quack, T., and Gool, L. V. (2013). Apparel classification with style. In *Computer Vision–ACCV 2012*, pages 321–335. Springer.

Chen, H., Gallagher, A., and Girod, B. (2012). Describing clothing by semantic attributes. In *Computer Vision–ECCV 2012*, pages 609–623. Springer.

Collin, C. and Wade, D. (1988). The barthel adl index: a standard measure of physical disability? *Disability & Rehabilitation*, 10(2):64–67.

Collin, C., Wade, D., Davies, S., and Horne, V. (1988). The barthel adl index: a reliability study. *Disability & Rehabilitation*, 10(2):61–63.

Fleury, A., Vacher, M., and Noury, N. (2010). Svm-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. *Information Technology in Biomedicine, IEEE Transactions on*, 14(2):274–283.

Funt, B., Cardei, V., and Barnard, K. (1996). Learning color constancy. In *IS&T/SID Fourth Color Imaging Conference*, pages 58–60.

Kalantidis, Y., Kennedy, L., and Li, L.-J. (2013). Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos. In *Proceedings of the 3rd conference on International conference on multimedia retrieval*, pages 105–112. ACM.

Liu, S., Feng, J., Song, Z., Zhang, T., Lu, H., Xu, C., and Yan, S. (2012a). Hi, magic closet, tell me what to wear! In *Proceedings of the 20th international conference on Multimedia*, pages 619–628. ACM.

Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., and Yan, S. (2012b). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3330–3337. IEEE.

Maitin-Shepard, J., Cusumano-Towner, M., Lei, J., and Abbeel, P. (2010). Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 2308–2315. IEEE.

Ramisa, A., Alenya, G., Moreno-Noguer, F., and Torras, C. (2012). Using depth and appearance features for informed robot grasping of highly wrinkled clothes. In

Robotics and Automation, International Conference on, pages 1703–1708. IEEE.

Self-maintenance, P. (1969). Assessment of older people: self-maintaining and instrumental activities of daily living.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124.

Stikic, M., Huynh, T., Laerhoven, K. V., and Schiele, B. (2008). Adl recognition based on the combination of rfid and accelerometer sensing. In *Pervasive Computing Technologies for Healthcare, 2008*, pages 258–263. IEEE.

Willimon, B., Birchfleld, S., and Walker, I. (2011). Classification of clothing using interactive perception. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1862–1868. IEEE.

Willimon, B., Walker, I., and Birchfield, S. (2013). A new approach to clothing classification using mid-level layers. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*.

Xia, L., Chen, C.-C., and Aggarwal, J. (2011). Human detection using depth information by kinect. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, pages 15–22. IEEE.

Zhang, Z. (2012). Microsoft kinect sensor and its effect. *Multimedia, IEEE*, 19(2):4–10.