Interacting with Digital Signage Using Hand Gestures

Qing Chen, François Malric, Yi Zhang†, Muhammad Abid,
Albino Cordeiro, Emil M. Petriu, and Nicolas D. Georganas⋆

DISCOVER Lab, University of Ottawa,
800 King Edward Avenue, Ottawa, ON K1N 6N5, Canada
†Institute of Computer Graphics and Image Processing,
Tianjin University, 300072 Tianjin, China
†yizhang@tju.edu.cn
{qchen,fmalric,mabid,acordeiro,petriu,georganas}@discover.uottawa.ca

Abstract. Digital signage is a very attractive medium for advertisement and general communications in public open spaces. In order to add interaction capabilities to digital signage displays, special considerations must be taken. For example, the signs' environment and placement might prevent direct access to conventional means of interaction, such as using a keyboard or a touch-sensitive screen. This paper describes a vision-based gesture recognition approach to interact with digital signage systems and discusses the issues faced by such systems. Using Haar-like features and the AdaBoosting algorithm, a set of hand gestures can be recognized in real-time and converted to gesture commands to control and manipulate the digital signage display. A demonstrative application using this gesture recognition interface is also depicted.

1 Introduction

Digital signage is a form of electronic display that is being used extensively to advertise targeted and impacting content to large audiences at public venues such as airports, shopping malls, and universities (see Fig. 1). Compared with traditional static signs, digital signage has the advantage of presenting dynamic multimedia digital content so that advertisers have more flexibility and scalability to adapt to different contexts and audiences with possibly less cost in the long run [1]. The digital signage industry is fast growing, and has been adopted by thousands of companies across many business sectors benefiting from the advantages it offers.

As illustrated in Fig. 2, a networked digital signage system includes a controller, a group of display panels and media players [2]. The controller uses the Internet to manage the whole system and deliver digital content to the display

⋆ Nicolas. D. Georganas holds a Cátedra de Excelencia at the Univ. Carlos III de Madrid and is a visiting researcher at IMDEA Networks, on leave from the School of Information Technology and Engineering, University of Ottawa.

Fig. 1. The digital signage systems are mounted at different public venues.

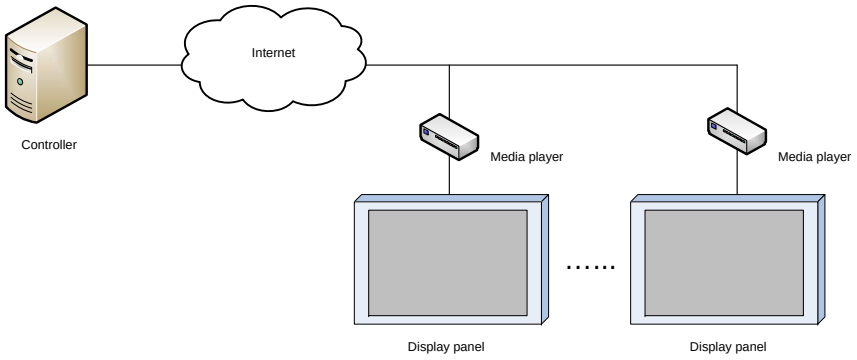


Fig. 2. The structure of a digital signage system.

panels. The display panels can be LED, LCD, plasma or other formats such as very large projected screens. Some indoor digital signs often include speakers and in some cases include digital cameras. Besides the controller and the display panel, a local media player, which is either a PC-based system or an embedded media player system, is needed as well. The local media player decodes data coming from the network so it can be rendered by the displays in the form of streamed digital content (e.g. video, flash animations, graphics, and other web content). The local media player can also store some content that is supposed to be played repetitively, and be remotely managed over the network by the controller to allow for content updates, schedule changes, and compliance reporting.

Digital signage is, in many ways, a mature technology. However, most digital signage systems are centrally controlled, and *interactive* digital signage is still in its infancy. Compared with nonreciprocal systems, interactive digital signage systems allow users to explore information according to their own interests so that more flexibility and involvement can be provided and users experience is greatly enriched. Currently, most interactive digital signage systems are based on touch screens which can capture users' finger movements on the screen surface and allow users to interact by touching pictures or words on the display. However, many prefer to install the displays in locations that are not at a dis-

tance directly accessible. To be visible by a wider audience and also to limit the risks of vandalism, they can be mounted high-up, or behind a display window. In these cases of digital signage where a touch screen is not feasible, touch-free interaction interfaces need to be explored. Some digital signage systems employ motion detection (e.g. infrared sensors) to trigger on and off of the display panel [3]. When a moving object is detected by the sensor, the digital signage system will turn on the display panel and boot up the system. If no moving object is detected in a pre-defined period of time, the display panel will go to sleep and the system enters standby until a moving object is detected again. The GestPoint® system developed by the GestureTek Company uses two cameras mounted overhead to recover the hand position and recognize the pointing gesture to control the display [4]. As only one gesture is recognized, this system can only implement a point and click function by using customized large icons and buttons on the display. Some digital signage systems provide keyboard and mouse for users to control the display. However, as most digital signs are installed in public venues, the lifespan of these peripheral devices are significantly shortened due to excessive usage.

To further enrich the user experience without the help of traditional human-computer interaction devices, hand gestures can be a good candidate for the touch-free interface between users and digital signage systems. Hand gestures represent a very expressive and a powerful human communication modality. As illustrated in Fig. 3, by attaching a webcam to the PC-based media player and utilizing advanced computer vision technologies, a vision-based hand gesture recognition system can track hand motions, identify hand gestures and convert them into direct control commands which digital signage systems can understand and respond to. With the gestural interface, users are able to control and ma-

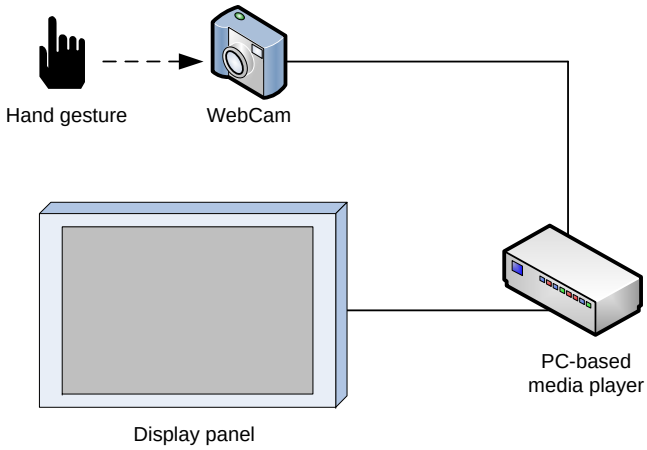


Fig. 3. The gesture-based interactive digital signage system.

nipulate the dynamic digital content by simply moving their hands and pointing at the information displayed on the screen.

In this paper, we will study vision-based hand tracking and gesture recognition for the purpose of interacting with digital signage systems. The hand tracking and gesture recognition will focus on bare hand tracking and recognizing gestures with a single webcam as the input device without help of any markers and gloves. The research will address both issues of hand tracking and gesture recognition in the context of interactive digital signage. Real-time signal processing algorithms will be used for tracking and identification of a set of hand gestures against different background and lighting conditions so that users can interact with the digital signage system in different environments. To demonstrate the effectiveness of the hand gesture interface, a prototype of gesture-based interactive digital signage system is implemented to enable the user to manipulate a web-based digital signage display.

2 Hand Tracking and Gesture Recognition

To use hand gestures to control digital signage systems, the gesture recognition system should meet the requirements in terms of accuracy, real-time performance and robustness. Recognition in a cluttered environment is considered a necessity since typical installations of digital signs can't guarantee a controlled environment. Popular image features to recognize hand gestures include skin color [5], [6], [7], [8], [9] and hand shape [10], [11], [12]. However, color-based gesture recognition algorithms face the challenge of eliminating objects with similar color such as a human arm and face. In order to solve this problem, users are often required to wear long-sleeve shirts and restrictions are imposed on the colors of other objects in the observed scene. Another problem of color-based algorithms is their sensitivity to lighting variations. When the lighting does not meet the specific requirement, color-based algorithms usually fail. For shape-based algorithms, global shape descriptors such as moments are used to represent different hand shapes. Most shape descriptors are pixel-based and the computation time is usually too long to implement a real-time system. Another disadvantage for shape-based approaches is the requirement for clean image segmentation, which is a difficult task for images with cluttered and noise-affected backgrounds.

To solve the problems faced by color and shape based algorithms, we employ a set of Haar-like features which have been used successfully for face detection [13]. As Fig. 4 shows, each Haar-like feature is a template of multiple connected black and white rectangles. The value of a Haar-like feature is the difference between the sums of the pixels' values within the black and white rectangular regions:

$$f(x) = W_{black} \cdot \sum_{black\_region} (pixel\ value) - W_{white} \cdot \sum_{white\_region} (pixel\ value)$$

where W_{black} and W_{white} are the weights that meet the compensation condition:

$$W_{black} \cdot black\_region = W_{white} \cdot white\_region$$

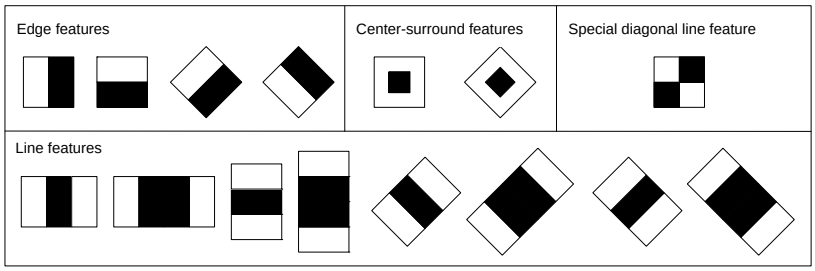


Fig. 4. A set of Haar-like features.

Different from image features such as skin color and hand shape, Haar-like features can encode ad-hoc domain knowledge which is difficult to catch using individual pixels. Haar-like features focus on the differences between the dark and bright areas within an image kernel. One typical example is that the eye region of the human face is darker than the nose region, and one Haar-like feature can effectively catch this property. Another advantage of Haar-like features is that they are more robust against noise and lighting variations due to their dependencies on the difference between the white and black rectangles. Noises and lighting variations affect the pixel values of the whole image kernel, and this influence can be effectively counteracted by the subtraction operation.

To detect the object of interest, a sub-window containing a specific Haar-like feature scan the whole image from its top-left corner to its bottom-right corner pixel by pixel. The object will be detected if the value of the Haar-like feature is above a certain threshold. To detect the object of different scales, the size of the sub-window needs to change accordingly. The sub-window starts from a very small initial kernel size and increase its width and height by multiplying a scale factor for the next scan. In this way, a number of sub-windows are discarded and the computation speed is improved. The bigger the scale factor, the faster the computation. However, the tradeoff is that the object with a size in between may be missed by the sub-window if the scale factor is too big.

A single Haar-like feature is certainly not enough to identify the object with a high accuracy. However, it is not difficult to find one Haar-like feature that has a slightly better accuracy than random guessing (i.e. accuracy better than 50%). In machine learning, these "better than random guessing" classifiers are called "weak classifiers". Boosting is a supervised machine learning algorithm to improve the overall accuracy stage by stage based on a series of weak classifiers [14]. A weak classifier is trained with a set of training samples at each step. This trained weak classifier is then added to the final classifier with a strength parameter proportional to the accuracy of this weak classifier. The training samples missed by the current weak classifier are re-weighted with a bigger value and the future weak classifier will attempt to fix the errors made by the current weak classifier so that the overall accuracy can be improved.

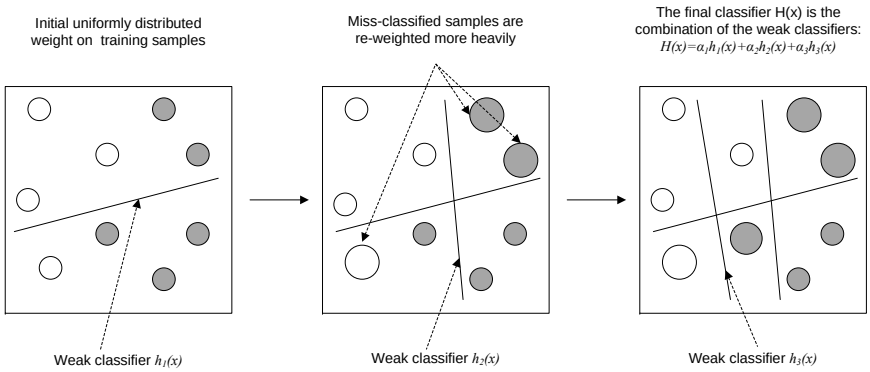


Fig. 5. The process of the AdaBoost learning algorithm.

The adaptive boost (AdaBoost) learning algorithm, which is first introduced by Freund and Schapire in [15], solved many practical difficulties of the earlier boosting algorithms (i.e. the first algorithm that could adapt to the weak learners). As illustrated in Fig. 5, the AdaBoost learning algorithm initially maintains a uniform distribution of weights over each training sample (in our case, the hand gesture images). It picks the Haar-like feature that yields the best classification accuracy in the first iteration. The weak classifier based on this Haar-like feature is added to the linear combination with a parameter proportional to its accuracy. In the second iteration, the training samples are re-weighted: training samples missed by the first weak classifier are boosted in importance so that the second Haar-like feature must pay more attention to these misclassified samples. To be selected, the second Haar-like feature must achieve a better accuracy for these misclassified training samples so that the overall error can be reduced. This iteration goes on by adding new weak classifiers to the linear combination until the required overall accuracy is met. The final training result is a strong classifier composed by a linear combination of the selected weak classifiers.

For the purpose of interacting with the digital signage system, we have selected four different hand gestures shown by Fig. 6. The selection of this gesture set is based on the consideration of easiness and naturalness for users to make these gestures. Furthermore, based on our experimental results, the selected gesture set proved able to avoid classification confusions possibly caused by the algorithm. 600 positive samples and 3500 negative samples are collected for each

(a) palm (b) fist (c) pointing (d) little finger

Fig. 6. The selected hand gestures for our system.

hand gesture. The positive samples are collected from 5 people with different hand shapes and skin colors. The numbers of the positive samples and negative samples are based on the experiment result: when the final classifier trained with 600 positive and 3500 negative samples already come close to the representation power, larger training sets do not affect the training result significantly. We set the overall false alarm rate at 1×10^{-6} to terminate the whole training process. For the "palm" gesture, a 12-stage classifier is trained with a hit rate at 98%. For the "fist", "pointing" and "little finger" gestures, the classifiers include 15 stages, 13 stages and 13 stages respectively. Their final hit rates are 97.7%, 99.7% and 99%. Fig. 7 shows some gesture recognition results for three different users with our trained classifiers. More detailed description of our approach is presented in [16].

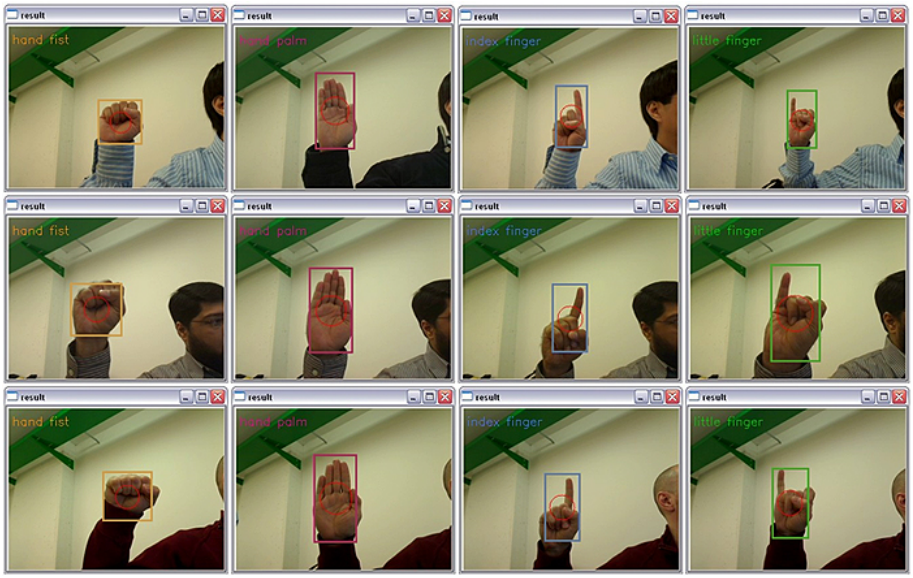


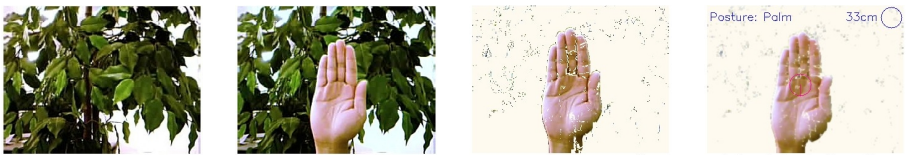
Fig. 7. Some gesture recognition results of the trained classifiers.

It is difficult to guarantee that every user would make the hand gestures in an ideal orientation. To evaluate the robustness of our classifiers against various hand rotations, we generate 500 test images with rotated hand gestures. The rotations include in-plane rotations and out-of-plane rotations. In-plane rotation means the image is rotated for a certain degree around "Z" axis perpendicular to the image plane. Out-of-plane rotations are the rotations around "X" axis or "Y" axis. According to our test results, for in-plane rotations, the detection rate decreases to 84% when the rotation degree reaches 10°. The detection rate reduces further to 62% when the rotation reaches 15°. For out-of-plane rota-

Fig. 8. The robustness of the "palm" classifier against different illuminations.

tions, the detection rates keep around 99% when the rotation reaches 20°. The detection rates reduce to 82% when the rotation reaches 30°.

Compared with color-based algorithms, one notable advantage brought by Haar-like features is the robustness against different lighting conditions, and consequently a certain degree of robustness to skin color differences. We tested our classifiers against images with different brightness values, Fig. 8 shows our test results. We tested the speed of the trained classifiers. The time required for each classifier to process one 320×240 true-color testing image is within 30 milliseconds. Adaptive background subtraction is used by our system to achieve the robustness against cluttered non-static backgrounds. Fig. 9 shows the background subtraction process. A 3×3 median filter and image dilation/erosion process are employed to reduce the noise produced by background subtraction. It is noticed the performance of the Haar-like features were improved after the noise removal measure is taken.



(a) background (b) "palm" gesture (c) bg. subtraction (d) noise removal

Fig. 9. The background subtraction and noise removal.

3 Interacting with the Digital Signage System

In this section, we introduce a gesture-based interactive digital signage system. The content of this digital signage, for demonstration purposes, is a web-based directory of the people in DISCOVER Lab at University of Ottawa (see Fig. 10). Each picture is linked to the individual's personal web page. A highlight blue box shows the active link. The available manipulations in this application include scrolling up/down the display, moving up/down the highlight box, opening/closing the individual's personal web page and zooming in/out the display. Instead of using static hand gestures to implement all of the manipulations, we have decided to integrate hand motions into our gesture command set so that

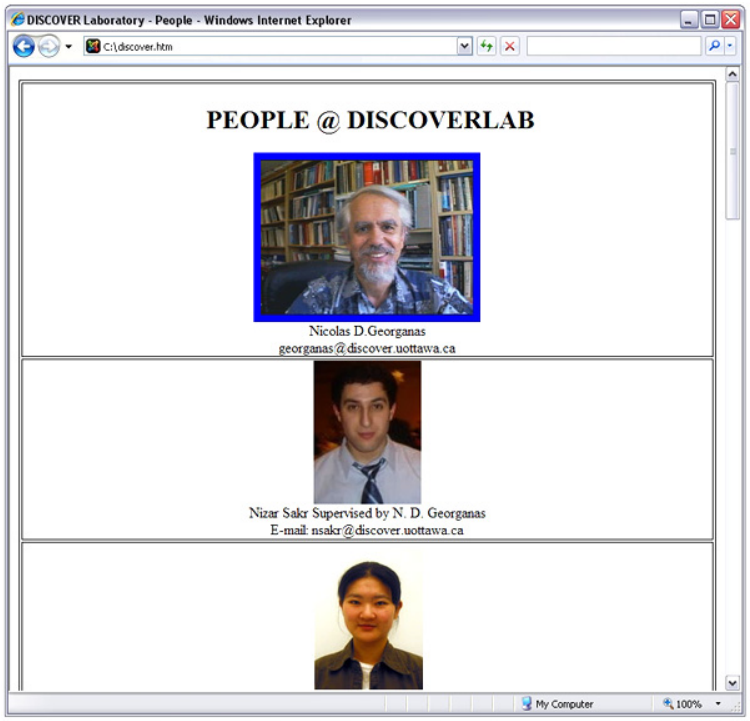
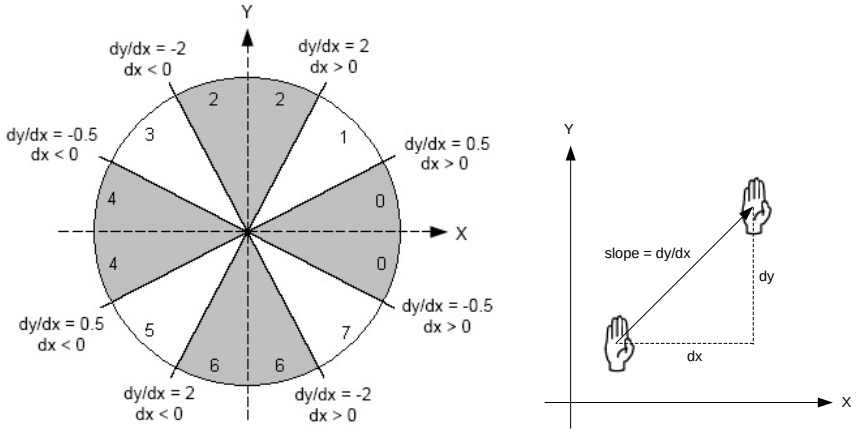


Fig. 10. The web-based contents of our digital signage system.

the user experience can be more intuitive and natural. For example, if the user wants to scroll up or scroll down the display, he simply moves his fist up and down. We use a set of direction primitives according to Fig. 11(a) to achieve this goal. The hand motion direction is estimated by computing the slope value based on the coordinates of the detected hand gestures in the image frame (see Fig. 11(b)).

A set of gesture commands are understood by integrating the recognized hand gestures and hand motions according to Table 1. These gesture commands consider the intuitiveness for the user to navigate the web-based content. To scroll up/down the display, the user simply moves his fist up/down accordingly in front of the camera. To move up/down the highlight box, the user just moves his palm up/down. If a particular person is interesting for the user, he first moves the highlight box to this person by moving his palm up/down, then he needs to perform the "point" gesture to open the individual's personal web page. To close the individual's web page, he just wags his little finger to go back to the main display. The user can also zoom in/out the display simply by move his fist back and forth. The display will be zoomed in if the size of the "fist" gets larger and vice versa. Fig. 12 shows a user interacting with the web-based digital signage system. A video clip of the demo can be found from [17].



(a) direction primitives (b) slope

Fig. 11. The assignment of direction primitives according to slope values.

Table 1. The initial Haar-like features selected by the trained classifiers.

Manipulations	Gesture commands
scroll up/down	2 / 6
highlight up/down	2 / 6
open web page	
close web page	
zoom in/out	back / forth

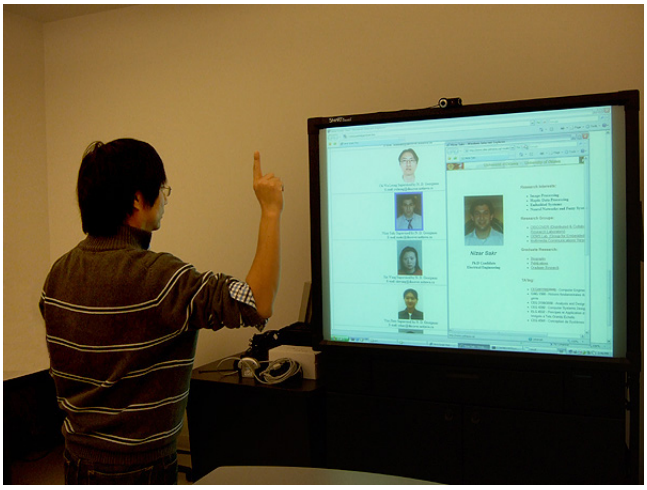


Fig. 12. A user interacts with the web-based digital signage prototype.

4 Conclusions

A gesture-based interactive digital signage system is introduced in this paper. This system implements a touch-free interaction interface using a set of gesture commands to manipulate the web-based content. With Haar-like features and the AdaBoosting algorithm, the system can track the hand and recognize a set of hand gestures accurately in real-time under different lighting conditions. Background subtraction and noise removal are used to achieve the robustness against cluttered backgrounds. With a webcam as the input device, by integrating the recognized gestures and hand motions, the user can control the digital signage display intuitively without the help of any other interaction devices.

For the future work, one improvement is to collect more diversified hand samples from different people for the training process so that the classifiers will be more user independent. Another improvement is to track and recognize multiple objects such as human faces, eye gaze and hand gestures at the same time. The relationships and interactions among these objects can be assigned with different meanings so that a richer command set can be integrated into a multiple user-based digital signage system. Moreover, other communication techniques such as voice recognition can also be integrated so that multimodal interactive capabilities can be achieved for a much richer user experience.

From a usability point of view, ongoing research is being done in order to determine a minimal set of intuitive gestures that can be robustly recognized by the system while enabling the user the execution of complex interactions. For this we are taking as a reference the set of user inputs acceptable by standard web browsers in internet navigation.

Acknowledgments: This project is supported by an NSERC Special Strategic Grant, by IBM Canada and by Larus Technologies.

References

1. Harrison, J. V., Andrusiewicz, A.: An emerging marketplace for digital advertising based on amalgamated digital signage networks. In: Proc. IEEE International Conference on E-Commerce, pp. 149–156 (2003)
2. Wang, P.: Digital signage 101: a quick introduction to those who are new to digital signage, http://digitalsignage.com/tools/articles
3. The DSE19M Economy Serie 19-Inch LCD Advertising Machine, http://www.industriallcd.com/d-dse19m-advertising.htm
4. GestPoint® Gesture Recognition for Presentation Systems, http://www.gesturetek.com/gestpoint/introduction.php
5. Wu, Y., Huang, T. S.: Non-stationary color tracking for vision-based human computer interaction. IEEE Trans. on Neural Networks, pp. 948–960 (2002)
6. Mckenna, S., Morrison, K.: A comparison of skin history and trajectory-based representation schemes for the recognition of user- specific gestures. Pattern Recognition, vol. 37, pp. 999-1009 (2004)
7. Bretzner, L., Laptev, I., Lindeberg, T.: Hand gesture recognition using multiscale colour features, hierarchical models and particle filtering. In: Proc. 5th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 405-410 (2002).
8. Imagawa, K., Matsuo, H., Taniguchi, R., Arita, D., Lu, S., Igi, S.: Recognition of local features for camera-based sign language recognition system. In: Proc. International Conference on Pattern Recognition, vol. 4, pp. 849-853 (2000)
9. Cui, Y., Weng, J.: Appearance-based hand sign recognition from intensity image sequences. Computer Vision Image Understanding, vol. 78, no. 2, pp. 157-176 (2000)
10. Ramamoorthy, A., Vaswani, N., Chaudhury, S., Banerjee, S.: Recognition of dynamic hand gestures. Pattern Recognition, vol. 36, pp. 2069-2081 (2003)
11. Ong, E., Bowden, R.: Detection and segmentation of hand shapes using boosted classifiers. In: Proc. IEEE 6th International Conference on Automatic Face and Gesture Recognition, pp. 889-894 (2004)
12. Ng, C. W., Ranganath, S.: Gesture recognition via pose classification. In: Proc. 15th International Conference on Pattern Recognition, vol. 3, pp. 699-704 (2000)
13. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 511-518 (2001)
14. Freund, Y., Schapire, R. E.: A short introduction to boosting. Journal of Japanese Society for Artificial Intelligence, vol. 14, no. 5, pp. 771-780 (1999)
15. Freund, Y., Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139 (1997)
16. Chen, Q., Georganas, N. D., Petriu, E. M.: Hand gesture recognition using Haar-like features and a stochastic context-free grammar. IEEE Transactions on Instrumentation and Measurement, vol. 57, no. 8, pp. 1562-1571 (2008).
17. Gesture-based interactive digital signage demo, DiscoverLab, University of Ottawa, http://www.discover.uottawa.ca/~qchen/my_presentations/gestureWeb.wmv