УДК 004.93'1

Вестник СибГАУ
2014. № 3(55). С. 162–167

АЛГОРИТМ РАСПОЗНАВАНИЯ ЖЕСТОВ РУКИ ЧЕЛОВЕКА НА ВИДЕОПОСЛЕДОВАТЕЛЬНОСТИ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ ДЛЯ РЕАЛИЗАЦИИ ИНТЕРФЕЙСОВ ЧЕЛОВЕКО-МАШИННОГО ВЗАИМОДЕЙСТВИЯ

М. М. Чудновский

Сибирский федеральный университет
Российская Федерация, 660028, г. Красноярск, просп. Свободный, 79
E-mail: chudnovskymax@gmail.com

Предложен алгоритм для распознавания жестов руки человека на видеопоследовательности в режиме реального времени, основанный на принципах цветовой кластеризации объектов, нахождения межэлементной разницы для видеопоследовательностей и контурного анализа. Решены задачи локализации, выделения ключевых признаков и распознавания жестов. Предложенный метод не требует длительного обучения классификаторов по массивам изображений жестов. Домен распознаваемых объектов задается параметрически на этапе первоначальной настройки алгоритма. Разработано специализированное программное обеспечение для экспериментального анализа ключевых показателей эффективности предложенного алгоритма. Показаны результаты оценки производительности метода и устойчивости распознавания. На основе полученных экспериментальных данных показана применимость алгоритма для распознавания жестов руки человека в режиме реального времени и возможность реализации на различных платформах вычислительной техники, что позволяет его использовать, прежде всего, в ракетно-космической отрасли для построения эффективных систем управления на основе жестовых интерфейсов.

Ключевые слова: жестовый интерфейс, распознавание жестов, цветовая кластеризация объектов, межэлементная разница видеопоследовательности, контурный анализ, инвариантные моменты.

Vestnik SibGAU
2014, No. 3(55), P. 162–167

A REAL-TIME ALGORITHM FOR HUMAN'S HAND GESTURE RECOGNITION ON VIDEO-SEQUENCE FOR HUMAN-COMPUTER INTERACTION INTEFACES

M. M. Chudnovsky

Siberian Federal University
79, Svobodny prosp., Krasnoyarsk, 660028, Russian Federation
E-mail: Chudnovskymax@gmail.com

The author presents the new real-time algorithm for hand gestures recognizing on a video sequence, based on the color clustering principles, interframe differences for video sequences and contour analysis principal. The first chapter of this paper contains the research of related work, which is based on Viola-Jones algorithm, the search for the skin color, wavelet transformation, localization of the centroid of the image, etc. The second chapter describes the new gesture recognition approach, and third includes experiment results. A proposed gesture recognition algorithm resolves two tasks: gesture localization and gesture recognition. The gesture localization task is resolved by compilation results of the image segmentation by human skin color clustering principle and motion search on the video sequences. The feature extraction task is resolved by invariant gesture contour moments analysis. A presented approach does not use any tutorial images. Gesture domain is not close and can be set up during initialization process. Key performance indicators (KPI) of the proposed approach are number of processed video sequence items per time unit and gesture recognition stability mark. The analysis of the performance of the proposed algorithm shows that the performance is at a high level. An average frame rate, which can function designed system is about 50 fps, which even exceeds the ability of the human visual perception. Recognition stability depends on environment condition changes, but the number of recognition errors can be reduced to a minimum by video device calibrating. The quality of recognition is 96%, which is a good indicator for such purpose algorithms. The KPI analysis, based on experimental data, is shown the applicability of the proposed approach in real-time gesture recognition systems, based on different hardware platforms. This fact allows to use the algorithm primarily in the aerospace industry to build effective management systems based on gestural interfaces.

Keywords: gestural interface, gesture recognition, objects color clustering, interframe difference for video sequences, contour analysis, invariant moments.

Введение. Одним из новейших методов человеко-машинного взаимодействия является использование систем распознавания определенных движений частей человеческого тела – жестов. Для «захвата» жестов человека могут использоваться различные устройства, например, ультразвуковые локаторы, кинематические датчики, системы структурированной подсветки и т. д., но наиболее распространенным устройством для получения данных о жестах пользователя является видеокамера. Исходя из этого, в работе рассматривается алгоритм распознавания жестов на видеопоследовательностях, так как подобный формат информации о жестах пользователя наиболее распространен ввиду распространенности устройств – видеокамер.

Задача распознавания жеста на видеопоследовательности заключает в себе несколько основных подзадач: локализацию жеста на элементе видеопоследовательности (поиск целевого объекта), выделение ключевых признаков локализованного жеста и сравнение этих признаков с эталонными значениями из базы данных. В зависимости от области применения жестового интерфейса указанные подзадачи могут решаться по-разному для обеспечения требований производительности и точности распознавания жестов.

Анализ существующих алгоритмов распознавания жестов. Как уже отмечалось, возможны различные подходы к решению подзадач распознавания жестов. Для локализации жестов на элементах видеопоследовательности часто используется метод Виолы–Джонса, например в работах [1, с. 103–105; 2]. Этот подход основан на принципах интегрального представления изображений, построения классификаторов на основе адаптивного усиления и каскадного комбинирования классификаторов. Данный метод требует наличия обучающей выборки изображений, обладает низкой скоростью обучения классификаторов, однако отличается высокой скоростью работы, что позволяет его использовать в системах реального времени. Важно отметить, что применимость метода Виолы–Джонса ограничивается конечным множеством возможных жестов для распознавания (в соответствии с обучающей выборкой). Алгоритм хорошо работает при небольшом угле отклонения, однако при отклонении угла более 30 градусов качество результатов резко падает. В качестве альтернативы может использоваться метод сегментации изображения на основе цветового кластера кожи [3, с. 2], который обладает низкой вычислительной стоимостью, однако не обеспечивает достаточной робастности жестов при наличии схожих по цвету объектов на исходной видеопоследовательности. Качество работы данного метода также сильно зависит от условий окружающей среды [4, с. 434].

Выделение ключевых признаков может осуществляться с помощью вельвет-преобразования, как показано в работах [1, с. 105–106; 2]. Этот подход показал совою эффективность во многих задачах обработки изображений. Важно отметить, что подобный подход также требует четко определенного домена возможных к распознаванию жестов и предварительного обучения. В работе [3, с. 2–3] используется принцип нахождения центроида локализованного объекта и расчета расстояния до максимально удаленных от цента точек. Данный подход инвариантен к положению жеста, однако сильно ограничен набором робастных жестов.

Алгоритм распознавания жестов. Классический сценарий использования жестовых интерфейсов предполагает захват и анализ движения человеческого тела на статичном фоне, без иных движущихся объектов. Исходя из данного предположения, а также в условиях работы в реальном времени для локализации жестов в данной работе предлагается комбинированный метод на основе сегментации исходного изображения с учетом цветового распределения кожи человека и нахождения межэлеметной разницы для видеопоследовательности – простейшего детектора движения. Предложенные методы обладают низкой вычислительной стоимостью, что позволяет построить высокопроизводительную систему. Приемлемое качество распознавания достигается за счет комбинации данных методов, при этом не требуется предварительное обучение, а домен возможных объектов для локализации не является замкнутым – можно задавать произвольные жесты, а также изменять их впоследствии.

Выделение ключевых признаков основано на принципах контурного анализа, что позволяет выделить характеристики, инвариантные масштабу и положению локализованного жеста. Подобный подход не требует предварительного обучения классификаторов, а значит, возможно динамически задавать классы при настройке алгоритма пользователем.

Сегментация изображения с учетом цветового кластера кожи человека. Принцип сегментации изображений на основе цветового кластера пикселей, образующих кожу человека, базируется на положении, что у людей различных рас составляющие цветового тона кожи меняются незначительно. Изменение состояния человека (эмоционального, физического) также слабо влияет на цвет его кожи [5, с. 2–3]. Результаты детектирования целевых объектов на основе цвета кожи не зависят от расположения и ракурса [6, с. 653–654]. Данные выводы позволяют говорить о том, что определение кожи по цвету (и сегментация изображений, основанная на этом принципе) является перспективным методом для применения в различных системах технического зрения.

Так как цвет кожи человека при контролируемом освещении занимает ограниченное подмножество цветового пространства [7, с. 593], эффективным и производительным решением является пороговый классификатор. Составленная в этом случае система неравенств для цветовых компонент классифицирует

элементы изображения по их соответствию цвету кожи человека. В зависимости от выбранных цветовых моделей неравенства будут различны, как и суммарная ошибка работы алгоритма выделения кожи [4, с. 433]. Данные анализа показывают преимущество ортогональной цветовой модели YCbCr [4, с. 432]. Границы цветового кластера кожи и показатель суммарной ошибки при контролируемом освещении для этой модели представлены в табл. 1.

Таблица 1

Показатели цветового кластера кожи

Цветовая модель	Граничные значения	Показатель общей ошибки детектирования
YCbCr	$25 \le Y < 220$ $100 \le Cb < 130$ $140 \le Cr < 190$	0,074

Количество операций при таком подходе соизмеримо с количеством элементов на изображении, а результат работы алгоритма представляет собой двоичную маску локализации. В качестве фильтров пред- и постобработки используются операции размытия и дилатации.

Нахождение межэлементной разницы для видеопоследовательности. Алгоритм вычисления межкадровой разности двух кадров для случая обработки цветного видео в формате RGB на вход принимает два элемента видеопоследовательности, представляющие собой две последовательности байт в формате одноканального изображения RGB. Оптимальным по производительности методом является приведение исходного полноцветного изображения RGB к изображению в серых тонах. В качестве первичного фильтра выступает операция эрозии исходного изображения. Далее производится вычисление попиксельных межкадровых разностей по следующей схеме:

$$D^*(i,j,n) = \left| D(i,j,0) - D(i,j,n) \right|, \qquad (1)$$

где $D(i,j,n)$ – значения яркости для текущего пикселя n-го кадра видеопоследовательности; $D(i,j,0)$ – значения яркости для текущего пикселя эталонного кадра фона; $D^*(i,j,n)$ – результирующее значение яркости для текущего пикселя n-го кадра видеопоследовательности. Полученное одноканальное изображение в результате межкадровой разницы подвергается пороговому преобразованию с заранее определенным параметром T:

$$M(i,j,n) = \begin{cases} 0, D^*(i,j,n) > T, \\ 255, D^*(i,j,n) < T, \end{cases} \qquad (2)$$

где $M(i,j,n)$ – значение элемента изображения после преобразования; T – пороговый уровень преобразования. Таким образом, на выходе алгоритма формируется двоичная маска. Детектор движения использует данные, предварительно сформированные системами обучения. В работе поиск движения базируется на данных о фоне сцены, на которой определяется перемещение целевых объектов – жестов руки человека. При получении нового кадра видеопоследовательности происходит подсчет абсолютной разницы между текущим кадром и данными об усредненном фоне. На основании этого получается разностная карта двух кадров, которая содержит информацию об объектах движения. Разностная карта (изображение в серых тонах) подвергается пороговому преобразованию, в результате чего получается бинарная маска локализации движения.

Сопоставление результатов локализации кожи и движения. Результатом определения кожи по цвету, а также разностного детектора движения являются изображения, содержащие так называемую бинарную маску, при наложении которой на исходный кадр можно получить объект, который детектируется в рамках рассматриваемого подхода. Исходя из того, что оба метода дают результат в одинаковом формате, а также учитывая тот факт, что оба метода используются для детектирования одного и того же целевого объекта, становится возможным объединение выходных изображений по принципу пересечения:

$$M_R = M_s \cap M_M, \qquad (3)$$

где M_s – бинарная маска, полученная в результате поиска кожи человека по цвету; M_m – бинарная маска, полученная в результате детектировании движения руки человека; M_R – результирующая маска. Результаты локализации жеста показаны на рис. 1.

а *б* *в*

Рис. 1. Бинарные маски локализации жеста: *а* – сегментация изображения с учетом цветового кластера кожи; *б* – межэлементная разность для видеопоследовательности (поиск движения); *в* – результирующая маска локализации жеста

Контурный анализ. Контурный анализ позволяет описывать, хранить, сравнивать и производить поиск объектов, представленных в виде своих внешних очертаний – контуров. Предполагается, что контур содержит всю необходимую информацию о форме объекта. Внутренние точки объекта во внимание не принимаются. Это ограничивает область применимости алгоритмов контурного анализа, но рассмотрение контуров позволяет перейти от двумерного пространства изображения к пространству контуров и тем самым снизить вычислительную и алгоритмическую сложность. Эффективным методом анализа и сравнения ограничивающих областей (контуров) является анализ их моментов.

Момент – это суммарная характеристика контура, полученная интегрированием (или суммированием) всех пикселей контура. Для функции $f(x,y)$, которая представляет собой изображение в серых тонах, моменты определяются следующим образом:

$$m_{p,q} = \iint x^p y^q f(x,y)dxdy, \qquad (4)$$

где p и q – порядок возведения в степень соответствующего параметра.

Использование моментов позволяет производить анализ и сравнение контуров. Однако характеристики контура, найденные по формуле (4), не пригодны для алгоритма распознавания, так как простые моменты являются зависимыми от площади локализованного жеста и ориентации его контура в системе координат. Для решения этих проблем необходимо рассчитать нормализованные центральные моменты:

$$v_{p,q} = \frac{\iint (x-x_c)^p (y-y_c)^q f(x,y)dxdy}{m_{00}^{\frac{p+q}{2}+1}}, \qquad (5)$$

где x_c, y_c – координаты центра тяжести изображения.

Нормализованные центральные моменты используются для получения характеристик контуров не зависящих от сдвигов и вращения, т. е. инвариантов, которые рассчитываются следующим образом [8, с. 183–184]:

$$\begin{aligned}
I_1 &= v_{20} + v_{02},\\
I_2 &= (v_{20} - v_{02})^2 + 4v_{11}^2,\\
I_3 &= (v_{30} - 3v_{12})^2 + (3v_{21} - v_{03})^2,\\
I_4 &= (v_{30} + v_{12})^2 + (v_{21} + v_{03})^2,\\
I_5 &= (v_{30} - 3v_{12})(v_{30} + v_{12})((v_{30} + v_{12})^2 -\\
&-3(v_{21} + v_{03})^2) + (3v_{21} - v_{03})(v_{21} + v_{03}) \times\\
&\times (3(v_{30} + v_{12})^2 - (v_{21} + v_{03})^2),\\
I_6 &= (v_{20} - v_{02})((v_{30} + v_{12})^2 - (v_{21} + v_{03})^2) +\\
&+ 4v_{11}(v_{30} + v_{12})(v_{21} + v_{03}),\\
I_7 &= (3v_{21} - v_{03})(v_{21} + v_{03})(3(v_{30} + v_{12})^2 -\\
&-(v_{21} + v_{03})^2) - (v_{30} - 3v_{12})(v_{21} + v_{02}) \times\\
&\times (3(v_{30} + v_{12})^2 - (v_{21} + v_{03})^2).
\end{aligned} \qquad (6)$$

Для каждого элемента видеопоследовательности выбирается контур локализованного жеста, после чего вычисляются его главные компоненты – инварианты I_1–I_7. Процесс распознавания заключается в сравнении главных компонент неизвестного контура с компонентами всех возможных жестов для распознавания, которые могут параметрически задаваться на этапе первичной настройки алгоритма (рис. 2).

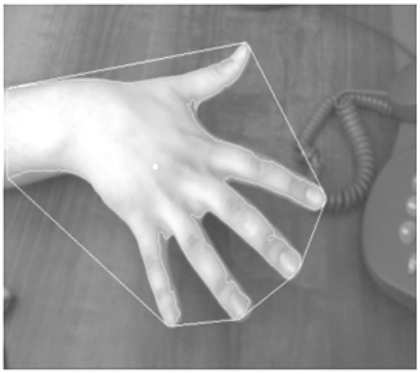

Рис. 2. Результат расчета контура локализованного жеста

Процесс распознавания жестов на видеопоследовательности. Работа алгоритма распознавания строится на двух основных этапах:

1) предварительная настройка;

2) распознавание жеста на элементе видеопоследовательности.

В свою очередь, этап распознавания состоит из локализации жеста на элементе видеопоследовательности, выделения главных компонент жеста и сравнения с заранее заданными (на этапе настройки) компонентами эталонных жестов. Алгоритмы этапов 1 и 2 представлены на рис. 3.

Экспериментальный анализ. Ключевыми показателями эффективности разработанного алгоритма являются производительность и устойчивость распознавания жестовой информации. Производительность системы оценивается таким параметром, как количество обработанных кадров в секунду, а устойчивость распознавания – показателем суммарной ошибки распознавания. При анализе качества распознавания, в связи с тем, что невозможно определить ошибку распознавания того или иного жеста (алгоритм работает на незамкнутом наборе жестов), для оценки качества работы системы используется показатель устойчивости распознавания изображения.

Для проведения экспериментов была разработана программа на платформе Microsoft.Net, реализующая предложенный алгоритм. При проведении экспериментов использовались внутренние индикаторы работоспособности разработанного ПО, а также использование сторонних программ для осуществления тестирования и мониторинга загрузки вычислительных мощностей аппаратной платформы. Эксперименты проводились на оборудовании компании HP усредненной конфигурации.

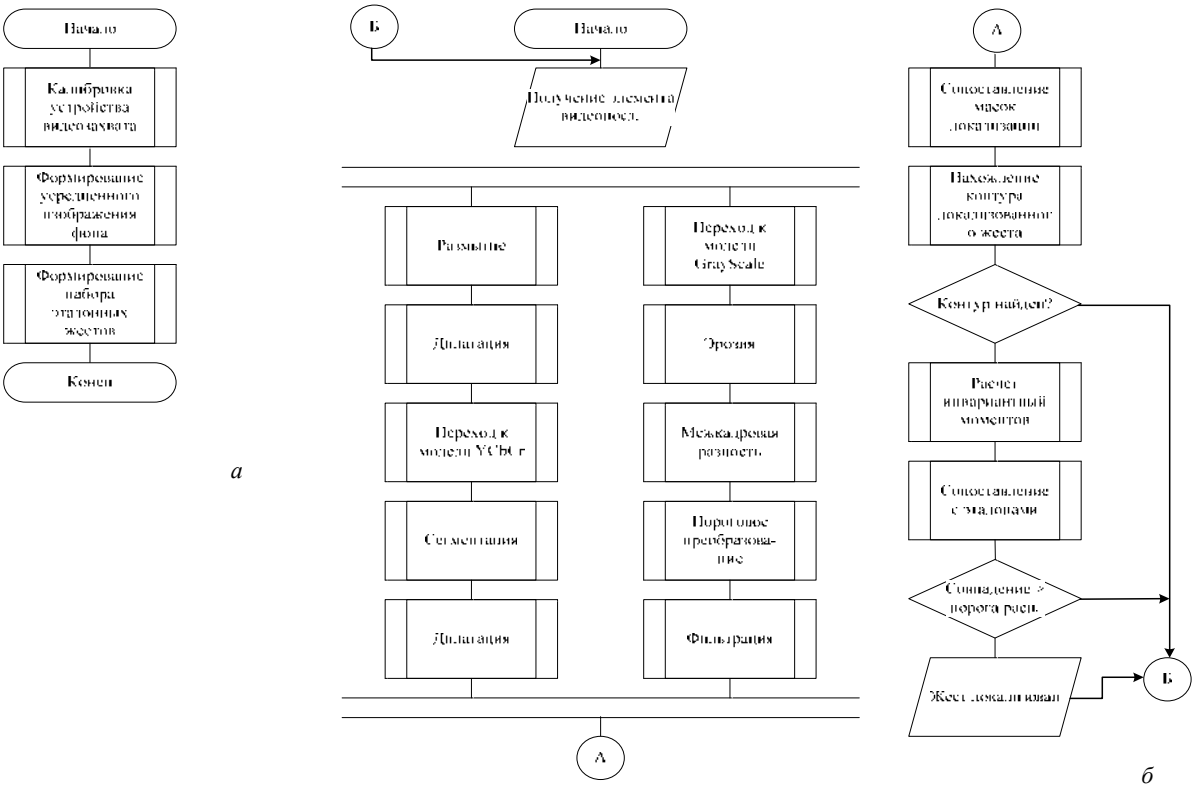

Рис. 3. Алгоритмы этапов процесса распознавания жеста: *а* – предварительная настройка; *б* – обработка элементов видеопоследовательности, процесс распознавания

Средняя производительность системы рассчитывается исходя из времени, которое требуется для обработки одной сцены. Количество кадров в секунду вычисляется как

$$FPS_{avg} = \frac{1}{T_{avg}}, \qquad (7)$$

где FPS_{avg} – максимально возможное количество кадров секунду; T_{avg} – время обработки одного кадра. На рис. 4 представлен график распределения времени обработки одного элемента видеопоследовательности.

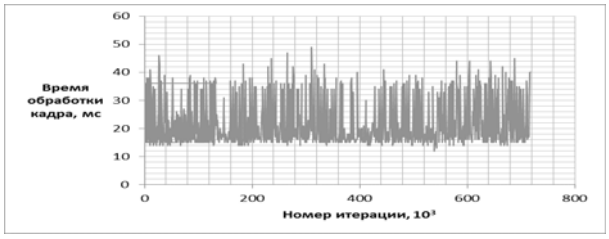

Рис. 4. Распределение времени обработки кадра

Анализ устойчивости распознавания основывается на покадровой оценке детектируемых контуров на сцене (анализ одного элемента видеопоследовательности представляет собой самостоятельный эксперимент), общая ошибка распознавания

$$Er = \frac{\left(Er_I \left|N - C\right| + Er_{II} \left|C\right|\right)}{\left|N\right|}, \qquad (8)$$

где Er_{I} – ошибка 1-го рода; Er_{II} ошибка 2-го рода; N – общее количество кадров; C – количество кадров, на которых изображен контрольный жест. Результаты исследования устойчивости распознавания представлены в табл. 2.

Таблица 2

Результаты оценки устойчивости распознавания

Вид ошибки устойчивости распознавания	Значение ошибки
Ошибка первого рода	0,0070 (0,7 %)
Ошибка второго рода	0,0740 (7,4 %)
Суммарная ошибка	0,0405 (4,05 %)

Анализ показателей эффективности предложенного алгоритма показывает, что производительность находится на высоком уровне. Средняя частота кадров, на которой может функционировать разработанная система, составляет около 50 fps, что даже превышает способности восприятия органов человеческого зрения.

Устойчивость распознавания зависит от изменений условий внешней среды, однако количество ошибок, возникающих при этом, снижается к минимуму посредством калибровки устройства видеозахвата.

Качество распознавания достигает 96 %, что является хорошим показателем для алгоритмов подобного назначения.

Новый алгоритм распознавания жестов руки человека, описанный в рамках работы, основывается на цветовой кластеризации кожи, поиске движения на видеопоследовательности и принципах контурного анализа. Предложенный алгоритм решает задачи локализации целевого объекта (кисти руки человека) на сцене, а также позволяет произвести анализ контекста его движения и формы – распознать жест. Анализ показателей эффективности алгоритма позволяет сделать вывод о возможном применении метода как в настольном, так в мобильном сегментах средств вычислительной техники. Широкий спектр применимости позволяет использовать алгоритм в различных направлениях науки и техники для построения эффективных интерфейсов человеко-машинного взаимодействия, прежде всего в ракетно-космической отрасли.

Библиографические ссылки

1. Фан Н. Х., Буй Т. Т., Спицын В. Г. Распознавание жестов на видеопоследовательности в режиме реального времени на основе применения метода Виолы–Джонса, алгоритма Camshift, вейвлет-преобразования и метода главных компонент // Управление, вычислительная техника и информатика : Вестн. Том. гос. ун-та.. 2013. № 2(23). С. 102–111.

2. Мурлин А. Г. [и др.]. Алгоритм и методы обнаружения и распознавания жестов руки на видео в режиме реального времени // Научный журнал КубГАУ [Электронный ресурс]. 2014. № 97 (03). URL: http://ej.kubagro.ru/2014/03/pdf/20.pdf (дата обращения 01.06.2014).

3. Malima A., Ozgur E., Cetin M. A Fast Algorithm for Vision-Based Hand Gesture Recognition for Robot Control // Proc. of the IEEE 14th. Antalya, 2006. P. 1–4.

4. Чудновский М. М., Русанова О. А. Исследование робастности принципа цветовой кластеризации объектов для построения систем человеко-машинного взаимодействия // Системный анализ и информационные технологии : V Междунар. конф. САИТ – 2013 (19–25 сент. 2013, г. Красноярск). В 2 т. Т. 2. Красноярск : ИВМ СО РАН, 2013. С. 431–435.

5. Shi L. Skin Colour Imaging That Is Insensitive to Lighting // Proc. of AIC Conf. on Colour Effects & Affects (June 15–18, Stockholm). 2008. no. 102.

6. Martinkauppi B. Detection of Skin Color under Changing Illumination: A Comparative Study // Proc. of the 12th Intern. Conf. on Image Analysis and Processing. (September 17–19, Mantova). 2003. P. 652–657.

7. Fleck M. Finding Naked People // Proc. of Fourth European Conference on Computer Vision (ECCV'96). (April 14–18, Cambridge). 1996. Vol. II. P. 592–602.

8. Hu M. K. Visual Pattern Recognition by Moment Invariants // IRE Trans. Info. Theory. 1962. Vol. IT-8. P. 179–187.

References

1. Fan N. Kh., Bui T. T., Spitsyn V. G. [Real-time Gesture recognition on video sequence based on Viola-Jones method, algorithm Camshift, the wavelet transform and principal component analysis]. *Vestnik TGU*. 2013, no. 2 (23), p. 102–111. (In Russ.)

2. Murlin A. G., Piotrovskiy D. L., Rudenko E. A., Yanaeva M. V. [Real-time algorithm and methods for hand gesture recognition on video]. *Nauchnyy zhurnal KubGAU*, 2014, no. 97 (03). (In Russ.) Available at http://ej.kubagro.ru/2014/03/pdf/20.pdf (accessed 01.06.2014).

3. Malima, A., Ozgur, E., Cetin, M. A Fast Algorithm for Vision-Based Hand Gesture Recognition for Robot Control. *Proc. of the IEEE 14th*. Antalya, 2006, pp. 1–4.

4. Chudnovskiy M. M., Rusanova O. A. [The object's color clustering principle research for human-computer interaction systems implementation] *Proc. 5th International conference "System analysis and information technologies"*. Krasnoyarsk, 2013, p. 431–435. (In Russ.)

5. Shi L. Skin Colour Imaging That Is Insensitive to Lighting. *Proc. of AIC Conference on Colour Effects & Affects,* June 15–18. Stockholm, 2008, p. 102.

6. Martinkauppi, B. Detection of Skin Color under Changing Illumination: A Comparative Study. *Proc. of the 12th International Conference on Image Analysis and Processing, September 17-19*. Mantova, 2003, p. 652–657.

7. Fleck M. Finding Naked People. *Proc. of Fourth European Conference on Computer Vision (ECCV'96),* Vol. II, April 14–18. Cambridge, 1996, p. 592–602.

8. Hu M. K. Visual Pattern Recognition by Moment Invariants. *IRE Trans. Info. Theory*. 1962, vol. IT-8, p. 179–187.