AUTOMATYKA • 2007 • Tom 11 • Zeszyt 3

Jaromir Przybyło*

Śledzenie cech charakterystycznych twarzy w systemie rozpoznawania mimiki**

1. Wprowadzenie

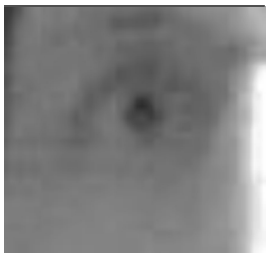
Obecnie obserwuje się zmianę podejścia w konstruowaniu interfejsów komputerowych w kierunku wykorzystania wielu równoległych sposobów komunikacji między użytkownikiem a maszyną. Podejście to często określane jest w literaturze jako interfejsy multimodalne (*multimodal interfaces*) [7]. Głównym sposobem komunikacji człowieka z komputerem jest – i zapewne długo jeszcze pozostanie – wprowadzanie danych przy pomocy klawiatury oraz interakcja z interfejsem graficznym systemu operacyjnego za pomocą myszki. Możliwości wprowadzania danych, rozszerzane są poprzez szereg urządzeń, bardziej specjalizowanych dla określonych zastosowań. W przypadku rozrywki (gry) klasycznych przykładem są różnego rodzaju manipulatory (joystick, trackball) bądź też gamepady. W zastosowaniach bardziej profesjonalnych (np. aplikacje typu CAD – *Computer Aided Design*) wykorzystywane są również ekrany dotykowe, tablety czy też manipulatory w postaci „rękawicy" pozwalającej na pracę z interfejsem 3D.

Obserwując komunikację człowieka z innymi ludźmi, łatwo można stwierdzić, że bardzo ważna jest tu także komunikacja niewerbalna, której istotnym elementem jest mimika twarzy. Ten kanał łączności między człowiekiem a komputerem i innymi systemami technicznymi (na przykład robotem medycznym albo wózkiem inwalidzkim) odgrywa szczególną rolę w przypadku niektórych osób dotkniętych głęboką niepełnosprawnością. Gdy kalectwo albo choroba odbiorą człowiekowi zręczność rąk konieczną do operowania myszką czy klawiaturą, oraz te same przyczyny utrudnią artykułowanie wyrazistych, nadających się do automatycznej interpretacji wypowiedzi słownych – mimika pozostaje jednym z ostatnich kanałów łączności chorego ze światem, w tym także ze światem systemów technicznych.

W takich sytuacjach alternatywę w sposobie sterowania komputerem, mogą stanowić metody i algorytmy rozpoznawania oraz analizy obrazów. W literaturze można znaleźć wiele rozwiązań dotyczących rozpoznawania mimiki na podstawie informacji wizyjnej [1, 4, 8]. Wspólną cechą powyższych rozwiązań jest konieczność lokalizacji charakterystycznych części twarzy, a następnie wykorzystania informacji o ich ruchu do sterowania.

* Katedra Automatyki, Akademia Górniczo-Hutnicza w Krakowie
** Pracę wykonano w ramach badań własnych (umowa AGH nr 10.10.120.39)

Niniejszy artykuł porusza zagadnienia lokalizacji oraz śledzenia cech charakterystycznych twarzy dla potrzeb wizyjnego interfejsu człowiek-komputer. Jako jedno z podstawowych założeń przeprowadzonych badań, przyjęto wykorzystanie popularnych kamer internetowych USB (*Universal Serial Bus*). Kamery takie charakteryzują się zazwyczaj ograniczonymi parametrami technicznymi. Typowa rozdzielczość otrzymywanego obrazu to 320×240 pikseli, natomiast liczba ramek na sekundę oscyluje w granicach 10 fps (*frames per second*). Problemem w przypadku takich kamer jest również niska jakość obrazu wynikająca z kompresji (rys. 1) lub wpływu układów automatycznej regulacji wzmocnienia [3]. W kontekście pracy człowieka z komputerem, uwzględnić należy również zmiany obrazu wywołane ruchami głowy człowieka oraz wpływ oświetlenia sceny.



Rys. 1. Wpływ kompresji sprzętowej kamery USB (widoczne kwadraty)

Praca posiada następujący układ. W rozdziale drugim przedstawiono algorytm lokalizacji cech z wykorzystaniem techniki dopasowania wzorców. Rozdział trzeci omawia zagadnienia odpowiedniego wyboru śledzonych cech. Omówienie wyników oraz ocena skuteczności lokalizacji cech zaprezentowane zostały w rozdziale czwartym. Artykuł kończy podsumowanie wniosków z przeprowadzonych badań.

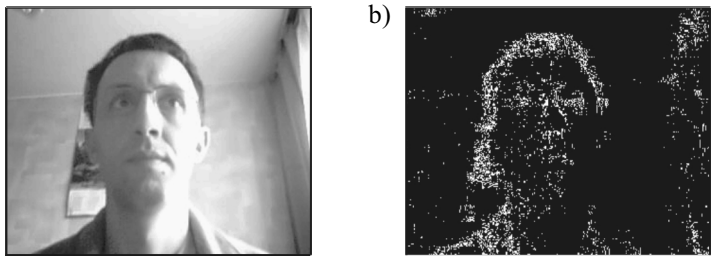
2. Lokalizacja cech charakterystycznych twarzy

Śledzenie obiektu (*visual tracking*) najczęściej definiowane jest jako lokalizacja obiektu na kolejnych klatkach sekwencji wideo. Lokalizacja taka realizowana może być poprzez binaryzację (z odpowiednio dobranym progiem) obrazu różnicowego lub odjęcie od aktualnej klatki obrazu tła. W wielu przypadkach powyższe techniki nie dają zadowalających rezultatów. Użycie obrazu różnicowego pozwala na lokalizację tylko obiektów poruszających się. Dodatkowym problemem są szumy (rys. 2), szczególnie widoczne na obrazach niskiej jakości otrzymywanych z popularnych kamer internetowych USB.

Z kolei odejmowanie obrazu tła jest nieskuteczne, w przypadku gdy zmieniają się warunki akwizycji (np. oświetlenie). Zastosowanie metod automatycznej generacji tła [9] poprawia w widoczny sposób rezultaty, nie zapewnia jednak rozróżniania między poszczególnymi obiektami.

Z tego powodu wymienione techniki stosuje się raczej do inicjalizacji właściwego algorytmu śledzenia cech twarzy [5], natomiast lokalizacja jest realizowana z wykorzystaniem metod dopasowania szablonów (*template matching*).

a) b)



Rys. 2. Przykładowa ramka obrazu (a) oraz moduł różnicy dwóch kolejnych ramek (b)

Metody dopasowania szablonów (inaczej wzorców) opierają się na wyszukiwaniu na obrazie (lub jego części) zadanego wzorca cechy według przyjętego kryterium dopasowania. Wyszukiwanie odbywa się poprzez wyznaczenie miary dopasowania szablonu dla każdego piksela obrazu (lub dla pikseli w określonym regionie poszukiwań (ROI – *region of interest*).

Wśród najczęściej stosowanych miar dopasowania wyróżnić możemy znormalizowaną korelację obrazu i szablonu (*normalized cross-correlation*) określoną następująco

$$C_{i,j} = \frac{\sum_m \sum_n \mathbf{I}_{m,n} \cdot \mathbf{T}_{m-i,n-j}}{\sqrt{\left[\sum_m \sum_n \mathbf{I}_{m,n}{}^2\right] \cdot \left[\sum_m \sum_n \mathbf{T}_{m-i,n-j}{}^2\right]}} \tag{1}$$

gdzie:

$C_{i,j}$ – wartość znormalizowanej korelacji w punkcie obrazu (i, j),

$\mathbf{I}$ – tablica pikseli obrazu o wymiarach $[Mi \times Ni]$,

$\mathbf{T}$ – tablica pikseli wzorca o wymiarach $[Mt \times Nt]$, $Mt < Mi$ oraz $NT < Ni$,

m, n – Indeksy: $0 = m = Mt$, $0 = n = Nt$.

Wartość funkcji znormalizowanej korelacji osiąga maksimum w miejscu największej zgodności wzorca z położonym pod nim fragmentem obrazu. Dokładniejsze omówienie stosowanych miar w technikach dopasowania wzorca można znaleźć w [10].

Przeprowadzone badania wykazały, iż w praktyce skuteczność lokalizacji metodą korelacyjną cech morfologicznych twarzy nie zawsze jest zadowalająca. Główne przyczyny to zmiany obrazu śledzonej cechy wywołane ruchami głowy człowieka (zniekształcenia perspektywiczne) lub zmianami oświetlenia sceny (rys. 7a). Na dokładność lokalizacji wpływ mają również zakłócenia wprowadzane m.in. przez elementy toru akwizycji obrazu (kamery internetowe USB).

Przykładowe wyniki lokalizacji wybranych cech przedstawione zostały w rozdziale 4.

3. Wybór cech obrazu

W kontekście wizyjnego systemu rozpoznawania mimiki, istotne jest zapewnienie dokładnej lokalizacji cech morfologicznych twarzy. „Zgubienie" śledzonej cechy, np. w przypadku ruchu głowy lub zmian oświetlenia, może spowodować błędną klasyfikację ruchu mimicznego, a w konsekwencji wygenerowanie innej akcji sterowania komputerem, niż było to intencją użytkownika systemu.

Dlatego istotnym etapem jest odpowiedni wybór cech oraz określenie wielkości wzorca. Tematyka ta poruszana jest w literaturze poświęconej śledzeniu cech obrazów (*feature tracking*) oraz problemom dopasowania obrazów (*image registration*). Jedną z częściej spotykanych technik jest wybór cech w miejscach gdzie istnieje zróżnicowana tekstura [6]. Takie założenie zwiększa prawdopodobieństwo poprawnej lokalizacji wzorca ponieważ obszary o bogatej teksturze zawierają więcej charakterystycznych elementów ułatwiających odszukanie wzorca.

Jedną z metod pomiaru zróżnicowania tekstury jest pomiar lokalnej entropii obrazu [2] określonej następującym wzorem

$$ent_{i,j} = -\sum_{k} p_k \cdot \log(p_k) \tag{2}$$

gdzie:

$ent_{i,j}$ – wartość lokalnej entropii dla obszaru ROI wokół piksela obrazu o współrzędnych (i, j),
p_k – znormalizowany histogram pikseli znajdujących się w ROI,
ROI – lokalne sąsiedztwo (ROI) analizowanego piksela obrazu o wymiarach $[w_h \times w_h]$,
k – liczba przedziałów histogramu.

Lokalną entropię wyznaczono dla każdego piksela obrazu, a następnie otrzymaną tablicę wartości (o wymiarach równych wymiarom obrazu) poddano operacji binaryzacji z progiem przyjętym doświadczalnie – 70% maksymalnej wartości entropii

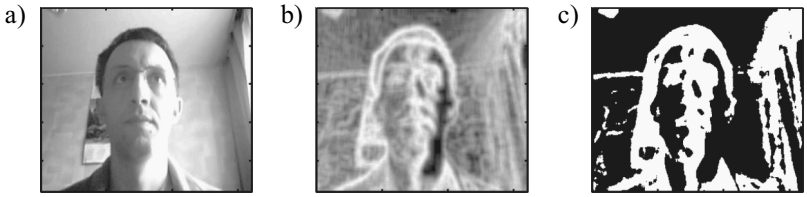
$$E_{i,j} = \begin{cases} 1 & \text{dla } ent_{i,j} \geq tresE \cdot \max(ent) \\ 0 & \text{dla } ent_{i,j} < tresE \cdot \max(ent) \end{cases} \tag{3}$$

gdzie:

$E_{i,j}$ – wartość lokalnej entropii obrazu po binaryzacji,
$tresE$ – wartość progu dla lokalnej entropii obrazu.

Piksele obrazu odpowiadające wartości 1 w tablicy E (nazywanej dalej mapą intensywności tekstury) odpowiadają miejscom o dużym zróżnicowaniu (rys. 3).

Przy wyborze cech pod uwagę należy również wziąć istnienie zakłóceń oraz szumów powstających w procesie akwizycji obrazu. Jest to szczególnie widoczne w przypadku wykorzystania kamer dających obraz o niskiej jakości (np. kamery internetowe USB). Drugą przyczyną zakłóceń są niewielkie ruchy głowy człowieka, w praktyce trudne do wyeliminowania bez stosowania specjalnych podpórek.



Rys. 3. Przykładowa ramka obrazu (a); lokalna entropia obrazu (b); mapa intensywności tekstury dla progu 70% (c)

Ilościowy pomiar szumów pozwala na wyeliminowanie z obrazu cech, które są silnie zakłócone. Pomiar szumów zrealizowano poprzez analizę zmienności wartości pikseli w czasie. Z sekwencji wideo wybrano kolejne ramki, dla których nie występuje ruch głowy, ruchy mimiczne lub zmiany oświetlenia sceny. Następnie dla historii każdego piksela wyznaczono wariancję, otrzymując tablicę o wymiarach takich samych jak analizowane ramki

$$s_{i,j} = \frac{1}{k-1} \cdot \sum_{k} (I_{i,j,k} - \bar{I}_{i,j})^2 \tag{4}$$

gdzie:

$s_{i,j}$ – wartość wariancji historii piksela obrazu o współrzędnych (i, j),

$I_{i,j,k}$ – wartość piksela obrazu (i, j) dla ramki k,

$\bar{I}_{i,j}$ – średnia wartość historii piksela (i, j),

k – liczba analizowanych ramek (długość historii piksela).

Wartości tablicy odpowiadają zmienności poszczególnych pikseli. Otrzymaną tablicę poddano operacji binaryzacji z progiem przyjętym doświadczalnie – 1,5% maksymalnej wartości zmienności (5). Piksele obrazu odpowiadające wartości 1 w tablicy S (nazywanej dalej mapą zmienności pikseli), odpowiadają miejscom o wysokim poziomie szumów (rys. 4).

$$ST_{i,j} = \begin{cases} 1 & \text{dla } s_{i,j} \geq tresS \cdot \max(s) \\ 0 & \text{dla } s_{i,j} < tresS \cdot \max(s) \end{cases} \tag{5}$$

gdzie:

$ST_{i,j}$ – wartość wariancji historii piksela (i, j) po binaryzacji,

$tresS$ – wartość progu dla wariancji obrazu.

Analizując mapę zmienności pikseli, można zauważyć, iż duży poziom szumów występuje na granicach obiektów (głowa i tło), w miejscach o dużej zmianie jasności oraz w pobliżu oczu. Zakłócenia w pobliżu oczu można wytłumaczyć mimowolnymi ruchami oczu oraz mruganiem.

a) b) c)

Rys. 4. Przykładowa ramka obrazu (a); wartość zmienność pikseli w czasie (b); mapa zmienności pikseli dla progu 1,5% (c)

Porównując otrzymane wyniki (mapa zmienności pikseli oraz mapa intensywności tekstury), można zauważyć, iż część pikseli o dużej intensywności tekstury charakteryzuje się jednocześnie dużym poziomem szumów. Uwzględniając powyższą obserwację, wybór cech zawężono do pikseli obrazu spełniających dwa następujące kryteria:

1) duże zróżnicowanie tekstury ($E_{i,j} = 1$),
2) małe szumy ($ST_{i,j} = 0$) .

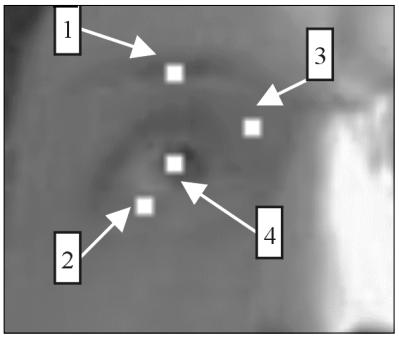
Przykładową mapę pikseli obrazu spełniających przyjęte kryteria przedstawia rysunek 5.

Rys. 5. Mapa cech obrazu spełniających przyjęte kryteria

4. Ocena skuteczności lokalizacji cech

W celu ewaluacji przyjętych kryteriów wyboru przeprowadzono porównanie lokalizacji cech dla czterech różnych przypadków – tj. cechy znajdującej się w miejscach obrazu (rys. 6):

1) o dużym zróżnicowaniu tekstury ($E_{i,j} = 1$) oraz małych szumach ($ST_{i,j} = 0$) –brew;
2) o małym zróżnicowaniu tekstury ($E_{i,j} = 0$) oraz dużych szumach ($ST_{i,j} = 1$) – dolna powieka;
3) o małym zróżnicowaniu tekstury ($E_{i,j} = 0$) oraz małych szumach ($ST_{i,j} = 0$) – wewnętrzny kącik oka (między powieką, nosem a brwią);
4) o dużym zróżnicowaniu tekstury ($E_{i,j} = 1$) oraz dużych szumach ($ST_{i,j} = 1$) – źrenica oka.



Rys. 6. Wybrane do śledzenia cechy twarzy: 1 – brew, 2 – dolna powieka, 3 – wewnętrzny kącik oka, 4 – źrenica oka

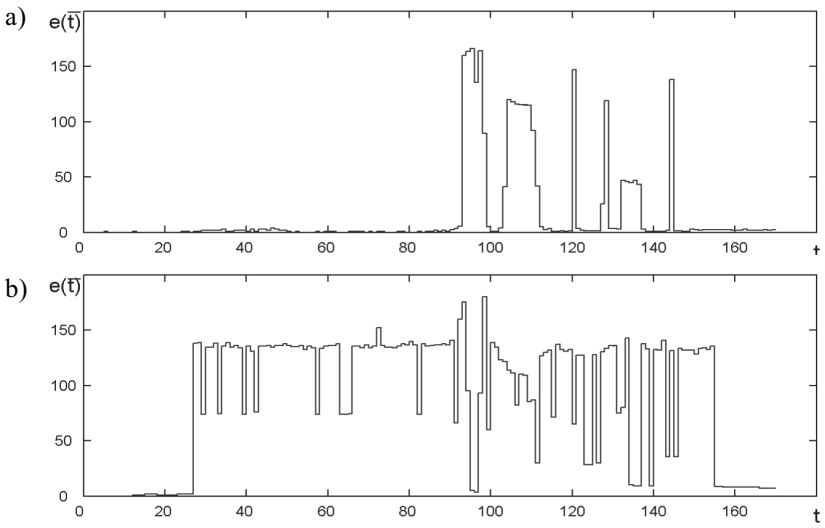
Testową sekwencję wideo pobrano za pomocą popularnej kamery internetowej (Creative Web Cam Pro) w rozdzielczości 320×240 pikseli, liczbę ramek na sekundę równą ~9, ramki obrazu w odcieniach szarości. Przyjęta liczba ramek na sekundę wynika z parametrów technicznych typowych kamer internetowych USB. Filmowana osoba poproszona została o wykonanie kilku typowych ruchów głową występujących podczas pracy z komputerem tj. spoglądanie lekko na boki oraz do góry i w dół, mruganie. W tabeli 1 zawarto opis sytuacji dla kolejnych ramek sekwencji.

Tabela 1
Opis akcji wykonywanych przez użytkownika dla kolejnych ramek testowej sekwencji wideo

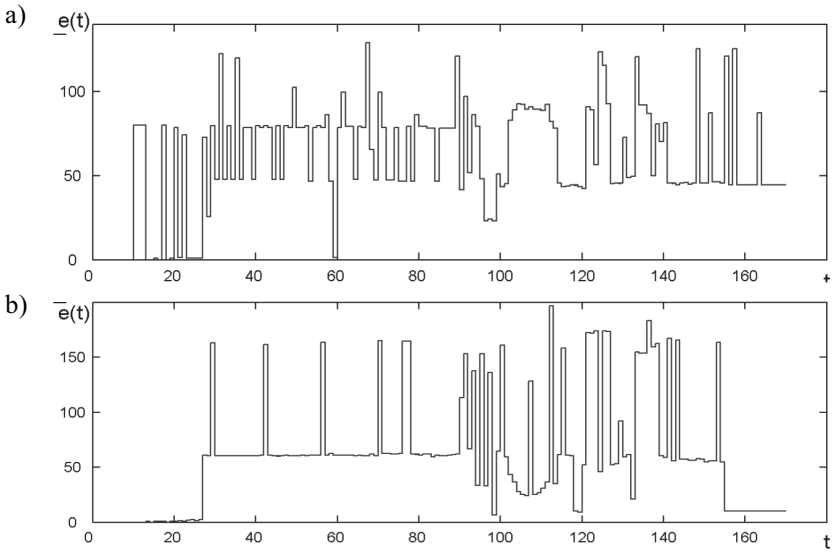
Numery ramek	Akcja
1–87	Głowa nieruchomo, mrugnięcia
88–101	Obrót głowy w prawo i z powrotem – dla ramek 93–98 zniekształcenia perspektywiczne w dużym stopniu utrudniają ręczną lokalizację cech
102–114	Obrót głowy w lewo i z powrotem – dla ramek 104–111 zniekształcenia perspektywiczne w dużym stopniu utrudniają ręczną lokalizację cech
117–128	Ruch głowy w górę i z powrotem
130–141	Ruch głowy w dół i z powrotem
41–42 49–51 62–68 77–82 149–153	Mrugnięcia oczami. Od ramki 27 zmiana punktu skupienia uwagi (ruch oka)

Jako wzorcowe położenie cech twarzy na kolejnych ramkach sekwencji, przyjęto lokalizację określoną przez człowieka za pomocą pomocniczej aplikacji, przy czym dla sytuacji przysłonięcia cechy (np. mrugnięcie) przyjmowano położenie relatywne do innych

elementów twarzy. Lokalizację przeprowadzono metodą korelacyjną dla wzorców cech o rozmiarach 13×13 pikseli. Wzorce zostały wskazane przez człowieka na pierwszej ramce sekwencji wideo. Na rysunkach 7 i 8 przedstawiono wyniki lokalizacji cech (poszczególne przypadki wymienione wyżej) na kolejnych ramkach sekwencji.



Rys. 7. Błąd średniokwadratowy lokalizacji cech: a) brew; b) dolna powieka



Rys. 8. Błąd średniokwadratowy lokalizacji cech: a) wewnętrzny kącik oka; b) źrenica oka

Dla cechy położonej w miejscu o dużym zróżnicowaniu tekstury oraz małych szumach (brew), błąd lokalizacji jest duży tylko dla momentów wychylenia głowy powodującego zniekształcenia perspektywiczne oraz częściowe przysłonięcie cech (rys. 7a). Wybór elementu twarzy zlokalizowanego w obszarze o małym zróżnicowaniu tekstury (wewnętrzny kącik oka) skutkuje bardzo szybkim „zgubieniem" śledzonej cechy (rys. 8a). Z kolei śledzenie cechy bogatej w teksturę, lecz położonej w punkcie gdzie występują duże zakłócenia (źrenica) powoduje, iż zmiany obrazu wywołane szumem lub mimowolną mimiką (mruganie) uniemożliwiają prawidłową lokalizację (rys. 8b).

Otrzymane wyniki potwierdzają tezę, iż przyjęte kryteria wyboru cech twarzy pozwalają zwiększyć skuteczność śledzenia.

5. Podsumowanie

Przeprowadzono badania skuteczności lokalizacji cech przy pomocy technik dopasowania szablonów (znormalizowana korelacja), z uwzględnieniem ograniczeń wynikających z parametrów technicznych popularnych kamer internetowych USB (niska rozdzielczość, duży poziom szumów) oraz typowych sytuacji występujących podczas korzystania z interfejsu przez człowieka.

Akcje wykonywane przez użytkownika podczas pracy z systemem komputerowym – takie jak ruchy głowy (np. spoglądanie na boki), oraz mimowolna mimika (np. mruganie) – wprowadzają duże zmiany w wyglądzie śledzonych elementów twarzy. Utrudnia to w dużym stopniu lokalizację przy pomocy techniki dopasowania szablonów.

Autor artykułu zaproponował kryteria wyboru cech, które pozwolą na zwiększenie skuteczności lokalizacji. Kryteria te opierają się na zawężeniu wyboru do miejsc obrazu, w których znajduje się zróżnicowana tekstura oraz niski poziom szumów. Opracowana metoda pomiaru szumów pozwala również na wyeliminowanie pikseli obrazu, które charakteryzują się dużą zmiennością wywołaną przez ruchy mimiczne twarzy. Przeprowadzona ocena skuteczności lokalizacji cech potwierdza poprawność przyjętych kryteriów wyboru.

Mimo zwiększenia skuteczności śledzenia cech, nadal występują sytuacje, dla których metoda dopasowania szablonów jest nieskuteczna. Przykładem mogą być znaczne ruchy głowy skutkujące bardzo dużą deformacją poszukiwanych cech (zniekształcenia perspektywiczne) lub wręcz ich przysłonięciem przez inne elementy twarzy. W takich sytuacjach wartość znormalizowanej korelacji jest większa dla innych elementów obrazu niż poszukiwana cecha. Objawia się to dużym błędem lokalizacji.

Rozwiązaniem powyższego problemu może być zawężenie rejonu poszukiwań do bezpośredniego otoczenia cechy oraz estymacja jej ruchu przy pomocy filtru Kalmana. Pod uwagę należy również wziąć konieczność adaptacji szablonu do zmieniających się warunków. Tematom tym poświęcony będzie inny artykuł.

Literatura

[1] Betke M., Gips J., Fleming P.: *The Camera Mouse: Visual Tracking of Body Features to Provide Computer Access For People with Severe Disabilities*. IEEE Transaction on Rehabilitation Engineering, 10(1), 2002, 1–10

[2] Gonzalez R.C., Woods R.E., Eddins S.L.: *Digital Image Processing Using MATLAB*. Chapter 11, New Jersey, Prentice Hall 2003
[3] Jabłoński M., Przybyło J., Wołoszyn P.: *Automatyczna segmentacja twarzy dla potrzeb interfejsu człowiek-komputer.* Półrocznik AGH Automatyka, t. 9, z. 3, 2005, 587–600
[4] Jilin Tu, Huang T., Hai Tao: *Face as Mouse Through Visual Face Tracking.* The 2nd Canadian Conference on Computer and Robot Vision (CRV'05), 2005, 339–346
[5] Kapoor A., Picard R.W.: *Real-Time, Fully Automatic Upper Facial Feature Tracking*. Proc. 5th Int. Conference on Automatic Face and Gesture Recognition, May 2002
[6] Kermad Ch., Collewet Ch.: *Improving Feature Tracking by Robust Points of interest Selection.* in 6th Int. Fall Workshop on Vision, Modeling and Visualization, VMV'2001
[7] Oviatt Sharon: *Multimodal Interfaces.* Chapter to appear in Handbook of Human-Computer Interaction, (ed. by J. Jacko & A. Sears), New Jersey, Lawrence Erlbaum 2002
[8] Przybyło J., Wołoszyn P., Jabłoński M.: *Rozpoznawanie jednostek czynnościowych mimiki twarzy na potrzeby interfejsu człowiek-komputer.* Półrocznik AGH Automatyka, t. 8, z. 3, 2004, 367–379
[9] Stauffer C., Grimson W.E.L.: *Adaptive background mixture models for real-time tracking.* Proc. IEEE CVPR, Fort Collins, Colorado, 1999, 246–252
[10] Vernon D.: *Machine Vision: Automated Visual Inspection and Robot Vision.* Chapter 6, Prentice Hall 1991