STUDIA INFORMATICA 2015
Volume 36 Number 1 (119)

Anna WÓJCICKA
Uniwersytet Pedagogiczny, Instytut Techniki
Uniwersytet Śląski, Instytut Informatyki
Roman SIMIŃSKI, Zygmunt WRÓBEL
Uniwersytet Śląski, Instytut Informatyki

DWUETAPOWA METODA EKSPLORACJI DANYCH POZYSKIWANYCH Z OBRAZÓW CYFROWYCH

Streszczenie. Głównym celem artykułu jest analiza problemu ekstrakcji wiedzy z obrazów cyfrowych oraz opracowanie spójnej metody integrującej algorytmy analizy obrazów ukierunkowanej na ekstrakcję cech jakościowych i ilościowych z metodami ekstrakcji wiedzy, wykorzystującymi dorobek dziedziny określanej mianem eksploracji danych oraz metod automatycznego wnioskowania.

Słowa kluczowe: systemy ekspertowe, analiza i przetwarzanie obrazów, odkrywanie wiedzy

THE TWO-STAGE METHOD THAT INTEGRATES IMAGE ANALYSIS AND DATA MINING

Summary. The main purpose of the paper is detailed analysis of the problem of knowledge extraction from images and to described a two-stage method which integrates digital image analysis focused on the extraction of quantitative and qualitative features and the knowledge extraction methods that use data mining analysis.

Keywords: system expert, image analysis and processing, data mining

1. Wstęp

Eksploracja danych jest aktualnie dziedziną badań naukowych dostarczającą znaczących rezultatów zarówno w ich aspekcie teoretycznym, jak i praktycznym. Opracowane metody eksploracji danych pozwalają na analizę zbiorów danych, ukierunkowaną na wykrywanie istotnych prawidłowości, zależności, powiązań i relacji, zapisanych niejawnie w owych zbio-

rach. Metody eksploracji danych zakładają najczęściej, iż danymi wejściowymi są odpowiednio przetworzone dane jakościowe oraz ilościowe, reprezentowane zwykle w postaci tabelarycznej. Rezultatem badań eksploracyjnych zwykle jest wiedza kodyfikowana z wykorzystaniem reprezentacji regułowej. Reguły będące wynikiem eksploracji wykorzystywane są w różny sposób, np.: do budowy klasyfikatorów lub baz wiedzy systemów zaliczanych do klasy systemów ekspertowych. Zarówno klasyfikatory, jak i systemy wykorzystujące bazy wiedzy oraz algorytmy wnioskowania są interesującym narzędziem badań w wielu dziedzinach, mają również znaczny potencjał wdrożeniowy, stanowiąc m.in. podstawę wielu użytecznych narzędzi wspomagania procesów diagnostycznych i decyzyjnych. Dla większości znaczących metod eksploracji zakłada się, że danymi wejściowymi będą wstępnie opracowane dane jakościowe i ilościowe, opisujące poddawane analizie informacje dziedzinowe. Taka forma danych wejściowych umożliwia stosunkowo łatwe pobranie danych z baz danych, arkuszy kalkulacyjnych, jak i odpowiednio sformatowanych plików tekstowych.

Jednocześnie można zaobserwować ciągły rozwój popularności metod pozyskiwania danych reprezentowanych w postaci obrazów. Spowodowane jest to ciągłym rozwojem technologicznym urządzeń do rejestrowania obrazów oraz wzrostem ich dostępności, do czego przyczyniają się m.in. ich spadające ceny. Dotyczy to zarówno urządzeń popularnego zastosowania, jak i specjalizowanych urządzeń, stosowanych w wielu dziedzinach badawczych, jak i zastosowaniach przemysłowych. Urządzenia te znajdują zastosowanie w wielu dziedzinach, stając się źródłem ogromnej ilości danych reprezentowanych w formie obrazów cyfrowych, niosących często wiele istotnych informacji. Jednak wydobycie tych informacji bywa utrudnione ze względu na wysoką liczebność przetwarzanych zbiorów obrazów oraz konieczność poddawania ich nietrywialnym analizom. Wydobywaniem informacji z obrazów cyfrowych skutecznie zajmuje się analiza obrazów, istnieje ogromna liczba uznanych publikacji wykazujących, że badania prowadzone w tej dziedzinie pozwalają na skuteczne odkrywanie cech ilościowych oraz jakościowych charakteryzujących obiekty występujące na zarejestrowanych obrazach [1, 2]. Wiele badań z zakresu analizy obrazów w istocie ma charakter eksploracyjny. Ukierunkowana jest właśnie na odkrywanie wyrażalnych jakościowo lub ilościowo informacji ukrytych w obrazach, a dotyczących zarejestrowanych na nich obiektów.

Wydaje się, że dane, będące wynikiem badań wykorzystujących analizę obrazów, powinny być idealnym materiałem wejściowym dla eksploracji danych. Analiza spójnej serii obrazów pewnych obiektów pozwolić może na wydobycie z nich serii interesujących cech owych obiektów, a poddanie ich dalszej analizie eksploracyjnej może pozwolić na odkrycie występujących regularności, uogólnień, powiązań i zależności. Te z kolei, zapisane w postaci np.: reguł, stanowić mogą ceny materiał wejściowy, pozwalający na stworzenie systemu z regułową bazą wiedzy, który z wykorzystaniem metod automatycznego wnioskowania pozwoli na

prowadzenie badań o charakterze klasyfikacyjnym lub będzie mógł wspomagać podejmowanie decyzji w zakresie określonej dziedziny problemu.

Przykładem dziedziny, w której powiązanie pomiędzy wynikami analizy obrazów a eksploracją danych może być bardzo istotne, jest medycyna. W diagnostyce medycznej powszechnie wykorzystuje się obrazy rejestrowane w trakcie różnorodnych badań, m.in.: rentgenowskich, ultrasonograficznych, tomograficznych, rezonansowych. Obrazy te są zwykle wykorzystywane do diagnostyki konkretnych pacjentów, jednak ich odpowiednio przygotowane archiwum może stanowić podstawę do przeprowadzenia ogólniejszych badań, poświęconych przykładowo analizie pewnej jednostki chorobowej. Analiza obrazów dostarczyć może istotnych informacji wydobytych z obrazów, eksploracja danych pozwolić może na wydobycie wiedzy zapisanej w tych informacjach. Wiedza sformułowana w postaci reguł pozwolić może zarówno na ukierunkowanie dalszych badań w danej dziedzinie, jak i ostatecznie na realizację konkretnego systemu wspomagania diagnostyki w ramach rozważanej jednostki chorobowej. Studia literaturowe wykazały, że podejście spójnie łączące te dwie dziedziny jest niezwykle rzadko stosowane. Oczywiście bardzo często konkluzje prac związanych z ekstrakcją danych z obrazów zawierają informację o planach kontynuacji badań z wykorzystaniem metod odkrywania wiedzy, jednak trudno znaleźć prace, dokumentujące badania, które w sposób uporządkowany i metodyczny łączą oba podejścia. Wskazać można wiele przyczyn takiego stanu rzeczy. Jedną z nich jest specjalizacja badań i zespołów badawczych, koncentrujących się albo na analizie obrazów, albo na eksploracji danych. Studia literaturowe oraz doświadczenie autorów zdobyte podczas dotychczasowych badań pozwalają na stwierdzenie, że istnieje brak spójnego powiązania, łączącego eksplorację cech w obrębie analizy obrazów z analizą eksploracyjną w ramach odkrywania wiedzy z danych. Oba te podejścia działają w sposób niezależny, posługując się odmiennym podejściem do problemu i stosowanymi metodami badawczymi. Kluczowym problemem wydaje się skoordynowanie celów i wyników ekstrakcji cech z celami oraz specyfiką metod eksploracji danych w taki sposób, aby wyniki działania metod analizy obrazu dostarczały właściwie opracowanych danych dla metod ekstrakcji wiedzy, zachowując korelację pomiędzy celami zakładanymi na każdym z etapów badań. Osiągnięcie proponowanej korelacji pomiędzy obiema metodami może nie być możliwe bez udziału ekspertów dziedzinowych. Zakłada się, że wiedza dziedzinowa może mieć kluczowe znaczenie dla powiązania obu podejść badawczych oraz weryfikacji poprawności i skuteczności tego powiązania.

Głównym celem praktycznym pracy jest opracowanie systemu informatycznego, wykorzystującego dwuetapową metodę eksploracji wiedzy z danych graficznych. System umożliwiać będzie ekstrakcję cech obiektów zapisanych na serii obrazów cyfrowych, co będzie realizowane z wykorzystaniem wyselekcjonowanych metod analizy obrazów. Cechy obiektów zapisane zostaną w formie wymaganej dla metod eksploracji danych, będą mogły zostać

poddane wyselekcjonowanym metodom eksploracji. Wynikiem dwuetapowego procesu będzie regułowa baza wiedzy. Na etapie korelacji obu etapów oraz na etapie weryfikacji wyników wykorzystana może być wiedza ekspertów dziedzinowych. Zakłada się jednak próbę opracowania metody działającej automatycznie. Proponowany system będzie posiadał również wbudowane algorytmy wnioskowania, pozwalające na realizację dedykowanych systemów wspomagania decyzji. Propozycja opracowania dwuetapowej metody eksploracji danych graficznych poprzedzona została badaniami wstępnymi, zrealizowanymi z wykorzystaniem obrazów z zakresu inżynierii materiałowej – złącza wytwarzane w technologii Friction Stir Welding oraz obrazów medycznych. Wyniki tych badań przedstawiono w publikacjach [3, 4].

Praca dokumentuje wczesny etap projektu, skupiając się na organizacji podjętego projektu badawczego i przedstawieniu wstępnych badań stanowiących studium wykonalności projektu.

2. Koncepcja organizacji badań

Rozdział ten prezentuje kolejne etapy podjętego problemu badawczego, precyzuje przyjęte metody analizy danych, przewidywane narzędzia oraz organizację procesu realizacji badań.

2.1. Problem badawczy

Aktualne badania proponowanej metody przeprowadzone zostały dla problemu wspomagania decyzji lekarza stomatologa w zakresie podjęcia odpowiedniego leczenia ubytków szkliwa, powstałych po zdjęciu aparatu ortodontycznego, w zależności od zastosowanych czynników odgrywających istotną rolę w procesie przyklejenia zamka ortodontycznego do zęba. W pierwszym etapie badań materiałem wejściowym do procesu analizy obrazów jest seria zdjęć cyfrowych zębów zarejestrowanych za pomocą optycznego tomografu. Przykład analizowanego obrazu został przedstawiony na rysunku 1, o rozmiarach 884x512 pikseli, zapisywanych w postaci plików z rozszerzeniem *.fda, *.fds. Wyodrębniono 910 obrazów zębów, na których zarejestrowano stan powierzchni po przeprowadzeniu kolejno pięciu etapów leczenia: przed ingerencją w szkliwo, polerowanie pastą ortodontyczną, wytrawienie i aplikacja systemu wiążącego przyklejenie zamka, usuniecie zamka, czyszczenie powierzchni zęba z systemu wiążącego.

Rys. 1. Powierzchnia zęba zarejestrowana za pomocą tomografu optycznego
Fig. 1. Registered tooth surface using an optical scanner

2.2. Organizacja prac badawczych

Proponowana metoda eksploracji danych z obrazów cyfrowych zakłada podział na dwa etapy. Etap pierwszy wykorzystuje wyselekcjonowane metody analizy obrazów, ukierunkowane na wydobywanie cech ilościowych i jakościowych obiektów przedstawionych na obrazach. Zakłada się, że danymi wejściowymi dla wybranych metod będą serie obrazów, przedstawiających analizowane obiekty. Na etapie ekstrakcji cech:

- ustalana będzie ich liczba, rodzaj, nazwy poszczególnych cech oraz nazwy lub zakresy wartości cech,
- ustalany będzie zestaw przekształceń graficznych, którym poddawane będą obrazy w celu ich standaryzacji oraz w celu uzyskania wymaganych cech.

Proces ten będzie realizowany z wykorzystaniem dedykowanego systemu informatycznego.

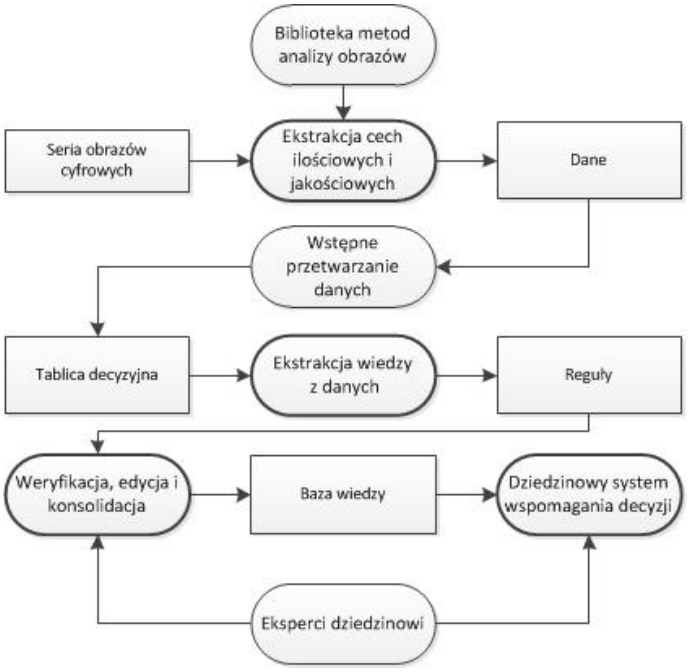
Wynikiem tego etapu są tabele z danymi, przyjmującymi postać systemu informacyjnego [5]. Zakłada się, że dane te poddawane będą przetwarzaniu wstępnemu, obejmującemu przetwarzanie danych brakujących, odstających, dyskretyzację wartości ciągłych. Wstępnie przetworzone dane utworzą tablicę decyzyjną [6], w której wskazany zostanie atrybut decyzyjny oraz atrybuty warunkowe.

Tablica decyzyjna stanowić będzie dane wejściowe dla drugiego etapu proponowanej metody. Etap ten obejmuje eksplorację danych, zakończoną wygenerowaniem reguł decyzyjnych. Proces ten poprzedzony będzie analizą spójności. Eksploracja danych opierać się bę-

dzie na podejściu bazującym na teorii zbiorów przybliżonych [5]. Wynikiem badań eksploracyjnych będą reguły decyzyjne. Zatem, na etapie ekstrakcji wiedzy:

- przeprowadzana będzie analiza spójności tablicy wejściowej, zakłada się wykorzystanie różnych metod usuwania niespójności,
- przeprowadzane będzie generowanie reguł decyzyjnych, zakłada się wykorzystanie różnych metod,
- przeprowadzana będzie analiza jakości uzyskanych reguł.

Drugi etap metody wykorzystywać będzie wsparcie ekspertów dziedzinowych, którzy z wykorzystaniem dedykowanego systemu będą mogli weryfikować uzyskane rezultaty z uwzględnieniem obrazów wejściowych. Szczególnym przypadkiem będzie analiza wyników ekstrakcji reguł dla różnych sekwencji obrazów, dotyczących tego samego problemu. Każda z sekwencji generować będzie bazę cząstkową. Proponowany system dostarczać będzie ekspertowi dziedzinowemu metod tworzenia wynikowej bazy wiedzy z wykorzystaniem baz cząstkowych. Schemat przedstawiony na rysunku 2 ilustruje koncepcję proponowanej metody.



Rys. 2. Schemat koncepcji proponowanej metody ekstrakcji wiedzy z obrazów cyfrowych
Fig. 2. Schema of the proposed methods for the extraction of knowledge from digital images

Ostatnim elementem przewidywanej metody będzie możliwość wykorzystania utworzonej bazy wiedzy do realizacji dziedzinowego systemu wspomagania decyzji. Polegać to będzie na wykorzystaniu wbudowanego w system modułu wnioskowania w przód i wstecz. Wnioskowanie to może być wykorzystane zarówno do badań eksperymentalnych i praktycznej weryfikacji otrzymanej bazy reguł, jak również do realizacji gotowego do wdrożenia systemu użytkowego.

2.3. Analiza obrazów stomatologicznych

W systemie wykorzystano obrazy otrzymane z tomografu optycznego przetworzone za pomocą algorytmu zaproponowanego w pracy [7]. Jest to m.in. kombinacja podstawowych algorytmów analizy i przetwarzania obrazu, tj.: filtracja, binaryzacja, konwersja, erozja, dylatacja, otwarcie i zamknięcie.

Powyższe operacje wykonano dla każdego zdjęcia, a wynikiem jest 910 macierzy o rozmiarze *512x128 pikseli* zawierających informację o grubości szkliwa dla poszczególnych etapów leczenia. Otrzymane dane pozwoliły na wyliczenie ubytków szkliwa powstałych po zakończeniu leczenia. Automatyzacja tego etapu polegała w pierwszym kroku na zbudowaniu systemu zarządzania i wczytania plików oraz na obsłudze automatycznego przetwarzania i analizie sekwencji obrazów opisanych w postaci danych liczbowych. Przyjęto, że nazwa każdego pliku zawierać powinna trzy człony: człon stanowiący numer badanego zęba, stały człon stanowiący numer procesu poprzedzonego znakiem „_”, oraz rozszerzenia .csv.

Opierając się na wiedzy ekspertów z zakresu analizy obrazów i lekarza stomatologa, do dalszej analizy został wybrany obszar zajmowany przez zamek ortodontyczny, co stanowiło obszar 100x100 centralnych pikseli (dla których odczytana została grubość szkliwa), czyli obszar 25122,25μm^2 powierzchni zęba. Pojedynczy piksel zajmował powierzchnię 5x11,7μm.

Kolejnym etapem badań było usystematyzowanie wiedzy na temat czynników odgrywających istotną rolę w procesie przyklejenia zamka do szkliwa, a wpływających na powstawanie ubytków szkliwa (rodzaj powierzchni zęba i jej właściwe przygotowanie, budowa podstawy przyklejanego zamka, rodzaj ortodontycznego materiału łączącego). To pozwoliło na wyodrębnienie informacji o charakterze danych opisujących proces przyklejania zamka ortodontycznego, jak i przygotowywania powierzchni zęba. Informacje te stanowiły podstawę do drugiego etapu analizy, ukierunkowanego na dobór metod przetwarzania danych, które pozwolą na odkrywanie i kodyfikację wiedzy o zależnościach pomiędzy czynnikami wpływającymi na ubytki szkliwa, jak i wielkością ubytków i doborem odpowiedniego leczenia.

2.4. Zestawienie i analiza otrzymanych baz reguł

Dane wejściowe poddano wstępnemu przetworzeniu, co pozwoliło wyodrębnić 182 obiekty, które opisano czterema atrybutami warunkowymi oraz atrybutem decyzyjnym – Damage. Następnie przystąpiono do etapu standaryzacji wartości atrybutu decyzyjnego opartego na wiedzy ekspertów – osiągnięto to poprzez wykorzystanie formuły matematycznej dostarczonej przez stomatologów, a pozwalającej określić wielkość ubytku szkliwa. W tym celu wyliczono średnie arytmetyczne, odchylenie standardowe oraz maksymalne i minimalne wartości grubości szkliwa dla każdego z przeprowadzonych etapów leczenia.

Proces dyskretyzacji atrybutu decyzyjnego przeprowadzono bazując na wiedzy ekspertów. W tym celu określono trzy przedziały wartości ubytków (small, medium, high), dla których powinno zostać podjęte leczenia. Po realizacji wstępnego przetwarzania danych w zbiorze 182 obiektów wykryto powtórzenia obiektów. Redukcja nadmiarowych obiektów wyodrębniła 60 unikatowych obiektów. Otrzymana tablica decyzyjna okazała się niespójna, dokonano usunięcia niespójności metodami ilościowymi, bazującymi na teorii zbiorów przybliżonych [5]. W przypadku równej liczby obiektów niespójnych, posiadających jednakową wartość atrybutu decyzyjnego, usunięcie niespójności zostało zrealizowane z udziałem ekspertów dziedzinowych.

W rezultacie przeprowadzonej analizy otrzymano 36 obiektów, które posłużyły do utworzenia reguł decyzyjnych. Wykorzystano metodę prostego generowania reguł dla każdego z obiektów tablicy. Reguły te pozwalają na określenie stopnia zniszczenia zęba w zależności od zastosowanych metod leczenia ortodontycznego.

Badania powyższe wykonano z wykorzystaniem autorskiego oprogramowania. Jednocześnie przeprowadzono badania z wykorzystaniem systemu RSES i metod eksploracji bazujących na teorii zbiorów przybliżonych. Z bazowego zbioru danych wyodrębniono 60 obiektów stanowiących zbiór uczący oraz liczący 28 obiektów zbiór testowy. W systemie RSES przeprowadzono eksperymenty polegające na wygenerowaniu reguł decyzyjnych oraz ich ocenie metodą *train and test*. Wykorzystano wszystkie dostępne w systemie RSES algorytmy generowania reguł, otrzymując odpowiednio: 22 reguły dla algorytmu wyczerpującego, 13 dla algorytmu genetycznego, 15 dla algorytmu pokryciowego, 15 dla algorytmu LEM2. Dla algorytmu wyczerpującego otrzymano pokrycie zbioru testującego o wartości 100%, dla algorytmu genetycznego 68% oraz dla algorytmu pokryciowego i LEM2 75%.

Osiągnięte rezultaty są weryfikowane przez stomatologów z Pomorskiego Uniwersytetu Medycznego w Szczecinie. Po jej uzyskaniu zostaną wznowione badania w zakresie generowania reguł, zostaną ponownie przeprowadzone eksperymenty algorytmami generowania reguł z wykorzystaniem systemów RSES, LERS, Weka i Rapid Miner.

3. Podsumowanie i wnioski końcowe

Głównym celem badawczym proponowanej pracy jest analiza problemu ekstrakcji wiedzy z obrazów oraz opracowanie dwuetapowej metody integrującej analizę obrazów cyfrowych ukierunkowaną na ekstrakcję cech jakościowych i ilościowych z metodami ekstrakcji wiedzy, wykorzystującymi eksploracyjną analizę danych. Proponowana metoda ma zmaksymalizować możliwości automatyzacji pozyskiwania wiedzy z obrazów, umożliwiając jednocześnie wykorzystanie wiedzy oraz kompetencji ekspertów dziedzinowych.

Analiza wyników dotychczasowych badań potwierdza realizowalność badań oraz osiągalność proponowanych celów. Wskazują one jednocześnie, że realizowane badania mają nietrywialny charakter.

Badania wstępne wskazują również na istotną wartość poznawczą proponowanych dalszych badań, przewidywany wkład w rozwój dziedziny, jaką jest eksploracja danych oraz potencjał aplikacyjny w wielu dziedzinach naukowych oraz zastosowaniach praktycznych. Głównym obszarem badań są do tej pory zastosowania medyczne. Jednak proponowana metoda, jak i realizowane w ramach pracy oprogramowanie będzie miało zastosowanie praktyczne wszędzie tam, gdzie zachodzi potrzeba ekstrakcji wiedzy z danych reprezentowanych w postaci obrazów cyfrowych.

BIBLIOGRAFIA

1. Bankman I.: Handbook of Medical Image Processing and Analysis. 2nd Edition. Academic Press, 2008, s. 1000.
2. Brennan D. J., Brandstedt J., Rexhepaj E., Foley M., Ponten F., Uhlen M., Gallagher W. M., O'Connor J. K., O'Herlihy C.: Tumour-specific HMG-CoAR is an independent predictor of recurrence free survival in epithelial ovarian cancer. BMC Cancer, 2010.
3. Wójcicka A., Simiński R., Wróbel Z.: Analiza metod predykcji parametrów zgrzewania metodą Friction Stir Welding. Studia Informatica, Vol. 35, No. 2 (116), Gliwice 2014.
4. Wójcicka A., Jędrusik P., Stolarz M., Kubina R., Wróbel Z.: Using analysis algorithms and image processing for quantitative description of colon cancer cells. Information Technologies in Biomedicine, Advances in Intelligent Systems and Computing, Vol. 283, 2014, s. 385÷395.
5. Pawlak Z.: Information System – theoretical foundation. WNT, Warszawa 1983.
6. Pawlak Z., Skowron A.: A rough set approach for decision rules generation. ICS Research Report 23/93, 1993.
7. Koprowski R., Machoy M., Woźniak K., Wróbel Z.: Automatic method of analysis of OCT images in the assessment of the tooth enamel surface after orthodontic treatment with fixed braces. Biomed. Eng. Online, Vol. 13, No. 48, 2014, s. 2÷18.

Abstract

The main objective of the research is a detailed analysis of the problem of knowledge extraction from images and to develop a two-step method of integrating digital image analysis

focused on the extraction of quantitative and qualitative features with the knowledge extraction methods that use data mining analysis. Analysis of the current state of knowledge in the analyzed problems leads to the conclusion that studies sign up in basic research, contributing to the development of the basic knowledge base. The project is a bridge between research in the field of image analysis and research devoted to knowledge discovery in data.

The system will allow the extraction of the characteristics of the objects saved on a series of digital images, which will be implemented with the use of selected methods of image analysis. Features of the objects are save in the form required for data mining methods, they can be subjected to specially chosen methods of exploration. The result will be a two-step process of rule-based knowledge base. The practical goal is to design and implement an IT system that will allow you to automate the selection of the parameters in dental treatment.

Adresy

Anna WÓJCICKA: Uniwersytet Pedagogiczny, Instytut Techniki, ul. Podchorążych 2, 30- 084 Kraków, Polska, awojcicka@up.krakow.pl; Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39, 41-200 Sosnowiec, Polska.

Roman SIMIŃSKI: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39, 41-200 Sosnowiec, Polska, roman.siminski@us.edu.pl.

Zygmunt WRÓBEL: Uniwersytet Śląski, Instytut Informatyki, ul. Będzińska 39, 41-200 Sosnowiec, Polska, zygmunt.wrobel@us.edu.pl