

TDK-dolgozat

Szerző: Szenkovits Annamária

Komponensek kinyerése képekből palimpszesztek digitális feldolgozására

Szerző:

Szenkovits Annamária, Babeş–Bolyai Tudományegyetem, Kolozsvár,
Matematika és Informatika Kar, informatika szak, alapképzés, III. év

Témavezető:

dr. Csató Lehel egyetemi docens, Babeş–Bolyai Tudományegyetem, Kolozsvár,
Matematika és Informatika Kar, Programozási Nyelvek és Módszerek Tanszék

Kivonat

A palimpszesztek olyan kéziratok, melyek eredetileg ókori írásokat tartalmaztak, ám nyersanyag hiányában a középkori szerzetesek ezeket a szövegeket levakarták, hogy helyükbe újat írthassanak. Nagyon sok ókori mű csak ilyen formában, töredékeiben maradt fenn, így igen fontos ezeket a kéziratokat tanulmányozni és megkísérelni a két írásréteg szétválasztását.

Dolgozatomban a számítógépes képfeldolgozásban használt módszerekkel vizsgálom a rendelkezésemre álló, palimpszesztekről készített képeket. A klasszikus képfeldolgozási műveletek (lineáris szűrők, gyakoriságelemzés) mellett elsősorban a természetfotók tanulmányozására jellemző statisztikai modelleket (PCA – Principal Component Analysis, ICA – Independent Component Analysis) próbálom ki a független komponensek keresésére.

Tartalomjegyzék

1. Bevezetés	3
1.1. Palimpszeszt	3
1.2. Palimpszesztek feldolgozásával kapcsolatos projektek	3
1.2.1. Hypereidoc	3
1.2.2. The Palimpsest Project	3
2. A függetlenkomponens-analízis módszer	4
2.1. A függetlenkomponens-analízis és a digitális jelekkelkülönítés	4
2.2. A függetlenkomponens-analízis alkalmazása releváns tulajdonságok kinyerésére . .	5
2.3. A függetlenkomponens-analízis matematikai háttere	5
2.3.1. Lapultság	6
2.3.2. Negentrópia	6
2.3.3. Mutuális információ	7
2.4. Előfeldolgozás	7
2.4.1. Az átlag normalizálása	7
2.4.2. Főkomponens-analízis és fehérítés	8
2.5. A FastICA algoritmus	8
3. A palimpszesztek és a függetlenkomponens-analízis	8
3.1. A kódtélpárti-jelenség és a palimpszesztek feldolgozása	9
3.2. A modell alkalmazása	9
4. További célkitűzések	11
5. Köszönetnyilvánítás	13
Hivatkozások	14

1. Bevezetés

1.1. Palimpszeszt

A palimpszeszt szó a latin 'palin' (újra), illetve 'psao' (vakarok) szavakból származik. A középkori szerzetesek, nyersanyag hiányában, a "régebbi", azaz ókori pergameneket hasznosították újra, mégpedig úgy, hogy a régi írást levakarták, kimosták, az új szöveget pedig ráírták a régre merőlegesen. Nagyon sok értékes ókori írás csak ilyen formában, azaz részeiben levakarva és fölülírva maradt fenn, így ezek megfejtése, feldolgozása igen fontos lehet. Rossz állapotuk miatt, illetve az ókori görög és latin nyelvezet miatt a feldolgozási folyamatnak számos akadály van.

Az egyik leghíresebb palimpszeszt az Archimédész-féle, mely az ókori tudós írásait tartalmazza. Ezt a tizedik században másolták az eredetiről, és a tizenharmadik században újrahasznosították: levakarták, 90°-kal elfordították és liturgikus szövegeket írtak rá. A palimpszeszt Archimédész írásai mellett az ókori görög szónok, Hypereides munkáját is, illetve Alexandros Aphrodisiensis kommentárjait is tartalmazza.

A digitális képfeldolgozás és a számítógépes karakterfelismerés segítséget nyújthat a bölcsészeknek a két írásréteg elkülönítésében, a lehetséges karakterek beazonosításában, illetve egy megfejtési javaslat kínálásában. Az ilyen fajta feldolgozás előnyei közé tartozik, hogy, lévén hogy csak képekkel dolgozik, nem ronsolja a törekeny kéziratot. Jelen dolgozat azt vizsgálja, hogy a két írásréteg milyen mértékben választható szét a különböző digitális képfeldolgozási módszerekkel, mint a függetlenkomponens-analízis (angolul Independent Component Analysis), illetve a főkomponens-analízis (Principal Component Analysis).

1.2. Palimpszesztek feldolgozásával kapcsolatos projektek

1.2.1. Hypereidoc

Dr. Horváth Zoltán vezetésével az ELTE Informatikai Karuk egy csapata 2008-ra fejlesztette ki a Hypereidoc szövegszerkesztőt ¹, mely a már megfejtett szövegek publikálását teszi lehetővé, a különböző konvencióknak, illetve jelöléseknek megfelelően.

1.2.2. The Palimpsest Project

A Palimpsest Project ² keretében egy kutatócsoport az Archimédész-palimpszeszt rekonstrukcióját, illetve oldalakra bontását végezték el (1. ábra). A csoport ezenkívül különféle optikai módszerek segítségével (LED, röntgen, ultraibolya, Multi-Spectral Imaging) a két réteg elkülönítésén is kísérletezett. Az így kapott eredményeket digitalizálták, olyan nagy felbontású, jó minőségű képeket

¹Hypereidoc szövegszerkesztő (<http://hypereidoc.elte.hu/?i=index/index>)

²A Palimpsest Project (<http://archimedespalimpsest.org>)



1. ábra. Palimpszesztlap rekonstrukciója és digitalizálása

nyerve, melyeken a két réteget két különböző színnel jelölték meg. A képek mindenki számára elérhetőek, ezekre alapoz e jelen munka is.³

A projekt keretében OCR-es feldolgozással is próbálkoztak⁴. 2004-ben a Rochester Institute of Technology képviselőjében Derek J. Walvoord kezdett neki egy, intelligens algoritmusokat alkalmazó rendszer fejlesztésének. A rendszer konkrét megvalósításáról a Palimpsest Project honlapja nem tartalmaz információkat.

2. A függetlenkomponens-analízis módszer

2.1. A függetlenkomponens-analízis és a digitális jelelkülönítés

A klasszikus jelelkülönítési feladat a koktélpárti-jelenség ([2], [3]). Képzeld el, hogy egy szoba két különböző pontjába egy-egy mikrofont szereltek fel; ugyanakkor, hogy a szobában két személy beszél egyszerre. A két mikrofon mindegyike rögzíti a két hangforrás valamilyen keverékét, attól függően, hogy melyiktől milyen távolságra van. Feladatunk, hogy kizárólag a mikrofonfelvételek ismeretében elkülönítsük a két eredeti hangforrást.

Jelöljük $s_1(t)$ és $s_2(t)$ -vel a két hangforrást, illetve $x_1(t)$ és $x_2(t)$ -vel a mikrofonok hangfelvételét, azaz a keverék felvételeket (mind a négy paraméter az idő, t függvényében változik). Az x változók felírhatóak az s -ek lineáris kombinációjaként, tehát a jelenség reprezentálható a következő egyenletrendszerrel (a t indexeket elhagyjuk):

$$\begin{cases} x_1 = a_{11}s_1 + a_{12}s_2 \\ x_2 = a_{21}s_1 + a_{22}s_2 \end{cases} \quad (1)$$

³A digitális palimpszeszt (<http://archimedespalimpsest.net/>)

⁴OCR-es feldolgozással kapcsolatos próbálkozások (http://www.archimedespalimpsest.org/pdf/archimedes_c.pdf)

Az a_{11} , a_{12} , a_{21} , a_{22} együtthatók értékét a mikrofonok és a hangforrások közti távolság határozza meg. Ha az együtthatók értékét ismernénk, az egyenletrendszer könnyen megoldhatnánk a klasszikus módszerekkel, esetünkben viszont mind az s , mind az a ismeretlen. A feladat bonyolultnak tűnik, viszont ha figyelembe vesszük, hogy az s -ek statisztikailag független valószínűségi változók (ha valamit tudunk az egyikről, az nem mond semmit a másikról, illetve fordítva), a megoldásnak egy jó közelítése adható. A függetlenkomponens-analízis módszere (angolul: Independent Component Analysis, ICA) az adatok függetlenségét kihasználva ad becslést az a együtthatókra, illetve különíti el az s forrásokat ([3]).

A függetlenkomponens-analízist eredetileg a koktéltarti-jelenséghez hasonló feladatok megoldására dolgozták ki, emellett viszont számos más érdekes alkalmazási területe létezik. Ilyen például az agy aktivitásának vizsgálata elektroencefalográfia segítségével.

2.2. A függetlenkomponens-analízis alkalmazása releváns tulajdonságok kinyerésére

A digitális jelfeldolgozó algoritmusok (pl. tömörítés, zajszűrés) bemenetét sokszor hatalmas méretű, magas redundanciájú adathalmaz (nagy adatmennyiség, mely kevés tényleges információt hordoz) képezik. A tulajdonságkinyerés módszerével a bemeneti adatokból olyan komponenseket nyerhetünk ki, melyek az adatok releváns jellemzőit tartalmazzák ([5]). Az adathalmaz reprezentálható ezen komponensek segítségével, csökkentve ezáltal a redundanciát, illetve a dimenziót.

Az adathalmaz releváns tulajdonságainak kinyerése lineáris transzformációs módszereken alapszik. A klasszikus módszerek közül leggyakrabban a Fourier-, Haar-, illetve koszinusztranszformációt alkalmazzák ([5]). Ezen elméleti modellek alkalmazásával ellentétben hasznosabb lehet a transzformációt a feldolgozandó adathalmaz alapján megbecsülni, és így a modellt az adott, sajátos esethez igazítani. A függetlenkomponens-analízis erre szintén kínál megoldási lehetőséget. Egy érdekes alkalmazása a természetfotók feldolgozása és ennek kapcsán a látókéregben lejátszódó adatfeldolgozás modellezése. A témakörrel a Helsinki Egyetem kutatócsoportja foglalkozott, Aapo Hyvärinen vezetésével. Eredményeiket a [2] könyvben foglalták össze. A könyv ugyanakkor kiindulópontot jelent saját munkámhoz is.

2.3. A függetlenkomponens-analízis matematikai háttere

A függetlenkomponens-analízis tárgyát képező feladatok általánosan az

$$x = A * s \quad (2)$$

alakban írhatóak fel, ahol az x és s oszlopvektorok, elemeik $x_1, x_2, \dots, x_n$, illetve $s_1, s_2, \dots, s_n$, az A "keverőmátrix" pedig az a_{ij} elemekből áll.

A függetlenkomponens-analízis generatív modell, azaz leírja, hogy a megfigyelt x adat hogyan

állítható elő, “keverhető ki” az s_i komponensekből. Az s_i független változók látens változók, nem tudjuk közvetlenül megfigyelni őket, mint ahogy az A mátrixot sem ismerjük. Kizárólag az x ismeretében kell megbecsülnünk az A -t és az s -t ([2], [3]).

A módszer arra alapoz, hogy a megfigyelt adat egymástól statisztikailag független, ugyanakkor “érdekes” (nem normál eloszlású) komponensek lineáris kombinációjaként írható fel. A függetlenség kifejezésére, illetve annak mérésére, hogy egy valószínűségi változó mennyire követi a normál eloszlást, többféle módszer létezik. A megfelelő függvények meghatározása után a feladat egy numerikus minimum-, illetve maximumkeresési feladattá válik.

2.3.1. Lapultság

Az s valószínűségi változó lapultsága vagy lapultsági mutatója azt fejezi ki, hogy a változó sűrűségfüggvénye hogyan viszonyul a normál eloszláshoz. A negyedik centrált momentum segítségével a következő formában írható fel:

$$kurt(s) = E\{s^4\} - 3(E\{s^2\})^2. \quad (3)$$

A lapultság értéke normál eloszlás esetében 0. Ha az érték pozitív, akkor az eloszlás sűrűségfüggvénye “csúcsosabb”, mint a normál eloszlásé. Ez egy 0 várható értékű, 1 szórású eloszlás esetében azt jelenti, hogy a sűrűségfüggvény értéke 0 körül, illetve a nagy értékeknél magas, míg a köztes értékeknél viszonylag alacsony. Negatív lapultság esetén a változó értéke 0 körül a konstanshoz közelít, míg nagy értékek esetén alacsony.

2.3.2. Negentrópia

Egy másik jó módszer annak mérésre, hogy egy valószínűségi változó hogyan viszonyul a normál eloszláshoz, a negentrópia. A negentrópia az információelméletből ismert entrópia fogalmára épül. Az entrópia gyakorlatilag azt jelenti, hogy mennyi információt nyújt egy valószínűségi változó megfigyelése. Minél véletlenszerűbb, előreláthatatlanabb, struktúrálatlanabb a változó, annál nagyobb az entrópiája. Az entrópia ugyanakkor szorosan összefügg az adott változó kódolási hosszával.

Egy diszkrét S valószínűségi változó esetében az entrópia a következőképpen értelmezett:

$$H(S) = - \sum_i P(S = a_i) \log P(S = a_i), \quad (4)$$

ahol a_i az S lehetséges értékei. Folytonos esetben, az $f(s)$ eloszlású s valószínűségi változó entrópiáját az alábbi képlet adja meg:

$$H(s) = - \int f(s) \log f(s) ds \quad (5)$$

Szintén információelméletből ismert, hogy az azonos szórású valószínűségi változók közül a normál eloszlásnak a legmagasabb az entrópiája. Az entrópia egy módosított változata a negentrópia:

$$J(s) = H(s_{gauss}) - H(s), \quad (6)$$

ahol s valószínűségi változó, s_{gauss} pedig az s szórásával megegyező szórású normál eloszlás. A negentrópia 0, ha s normál eloszlású, különben egy pozitív érték.

A negentrópia tehát szintén használható a normalitás becslésére. Hátránya, hogy viszonylag nehéz számításokat végezni vele, így a gyakorlatban csak közelítésekkel dolgoznak.

2.3.3. Mutuális információ

Két valószínűségi változó egymáshoz viszonyított kapcsolatát a – szintén információelméletből ismeretes – mutuális információval vizsgálhatjuk. A mennyiség a változók függetlenségét fejezi ki. Gyakorlatilag egyenértékű a változók együttes eloszlásának, illetve a marginálisok szorzatának Kullback-Leibler-távolságával.

Egy $p(x)$, illetve $q(x)$ eloszlásfüggvényű valószínűségi változó Kullback-Leibler-távolsága (vagy relatív entrópiája) az alábbi összefüggéssel adható meg:

$$D_{KL}(X_p||X_q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}. \quad (7)$$

A két változó egyfajta távolságát fejezi ki, és értéke akkor és csak akkor 0, ha $p = q$.

A fentiek tükrében tehát egy adott X és Y valószínűségi változó esetében, ahol az együttes eloszlás $p(x, y)$, a marginális eloszlások pedig $p(x)$ és $p(y)$, a két változó mutuális információját a következő:

$$I(X; Y) = D_{KL}(p(x, y)||p(x)p(y)). \quad (8)$$

Értéke 0 akkor és csak akkor, ha a két változó statisztikailag független, különben pozitív.

2.4. Előfeldolgozás

A függetlenkomponens-analízis alkalmazása előtt hasznos lehet az adatokon előfeldolgozást végezni. Ez megkönnyíti a számításokat és a komponensek különválasztását ([2], [3]).

2.4.1. Az átlag normalizálása

A legalapvetőbb művelet amit előfeldolgozásként végezhetünk az, hogy x -ből kivonjuk a várható értéket ($m = E\{x\}$), így x -ből egy 0, azaz normalizált várható értékű valószínűségi változó válik.

Ebből az következik, hogy az s -et is ilyen formában kapjuk meg, ám az A mátrix megbecsülése után s átlagát megkaphatjuk az $A^{-1}m$ alakban, ahol m az előzőleg kivont átlag.

A módszer a DC-komponens eltávolításaként elsősorban a villamosmérnöki eljárásokban használt. A DC-komponensre úgy tekintenek, mint redundáns, tehát “érdekes” információt nem hordozó tényező, ezért a számítások leegyszerűsítése végett előnyös eltávolítani.

2.4.2. Főkomponens-analízis és fehérítés

További előfeldolgozásként alkalmazhatjuk a bemenetre a főkomponens-analízist (angolul Principal Component Analysis – PCA) az adatok fehérítésére. Ez lényegében azt jelenti, hogy x -en olyan lineáris transzformációt hajtunk végre, mely során egy fehérített $\tilde{x}$ vektort kapunk, azaz komponensei nem korreláltak, szórásuk pedig 1. Más szóval: az $\tilde{x}$ kovariancia-mátrixa megegyezik az identitás mátrixszal:

$$E\{\tilde{x}\tilde{x}'\} = I. \quad (9)$$

A fehérítési transzformáció mindig létezik. A leggyakrabban használt módszer a meghatározására a korrelációs mátrix sajátérték-dekompozíciója. A függetlenkomponens-analízis leegyszerűsödik fehérítés során, ugyanis a megbecsülendő paraméterek dimenziója csökken, ugyanakkor, az eredeti A keverőmátrix helyett egy $\tilde{A}$ ortogonális mátrixot kell meghatároznunk ([2], [3], [4]).

2.5. A FastICA algoritmus

A függetlenkomponens-analízist megvalósító népszerű és elterjedt algoritmus a Helsinki Egyetem kutatója, Aapo Hyvärinen által kidolgozott ingyenes FastICA MATLAB-csomagban található⁵ [1]. Az algoritmus az eloszlás normalitásának mérésére a negentrópia közelítését használja, a negentrópia maximalizálására pedig fixpont-iterációt alkalmaz.

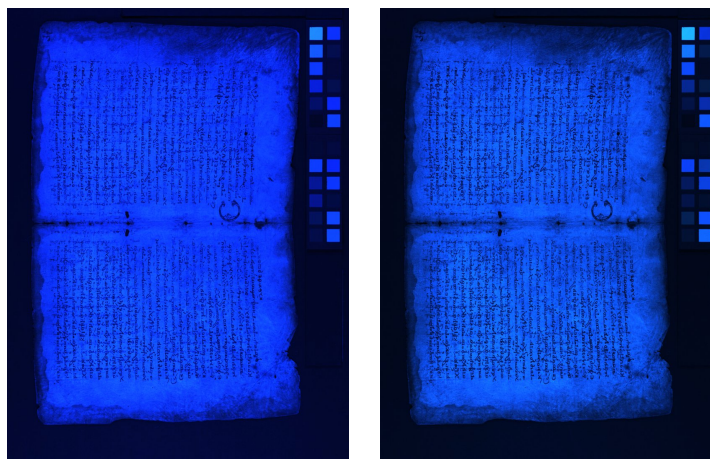
A fastica eljárást a $[icasig, W, A] = fastica(X)$; alakban hívhatjuk meg. X a mintákat tartalmazó mátrix; icasig a becslés szerint kapott független komponenseket tartalmazza, dimenziója megegyezik az X dimenziójával; A a transzformációs vagy keverőmátrix, W ennek inverze. Megjegyzendő, hogy a fastica algoritmus a W -t közelíti, majd a W pszeudo inverze alapján kapja meg az A -t.

Munkám során magam is a fent említett FastICA csomagot használok.

3. A palimpszesztek és a függetlenkomponens-analízis

Ismétlésképp: a palimpszesztek olyan kéziratok, melyek két különböző korból származó szövegeket tartalmaznak. A középkori szerzetesek, nyersanyag szűkében az eredeti, ókori írásokat levakarták a könyvtáraikban levő pergamenekről, hogy újabb szövegeket másoljanak rájuk.

⁵A csomag, illetve a hozzá tartozó dokumentáció elérhető itt: (<http://research.ics.tkk.fi/ica/fastica/>)



2. ábra. Palimpszesztlap két különböző fényű megvilágításban

Az Archimedes Project keretében a stanfordi kutatócsoport a palimpszeszt oldalainak elkülönítése után egy-egy oldalról több felvételt is készített. A kéziratokat megvilágították különböző fényű LED-diódákkal, röntgensugarakkal, illetve ultraibolya sugarakkal. Az 2. ábra egy palimpszesztlap két különböző fényű megvilágítását szemlélteti.

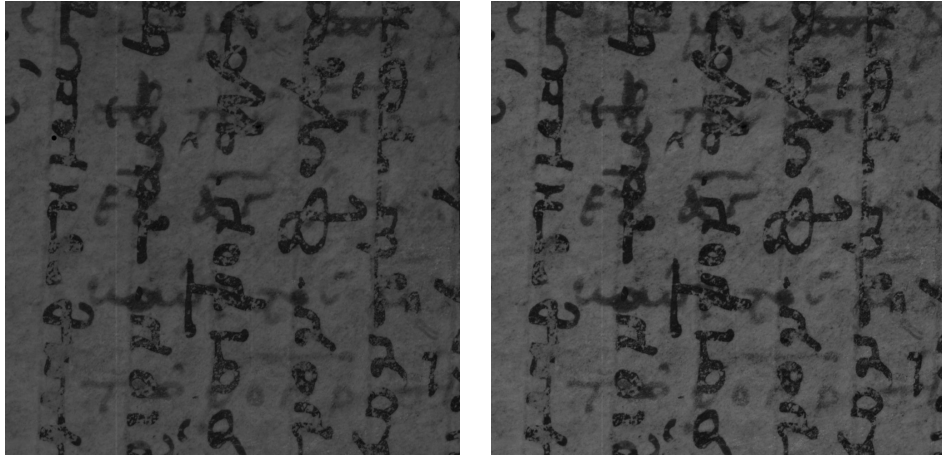
3.1. A koktélparti-jelenség és a palimpszesztek feldolgozása

Ha azt vizsgáljuk, milyen módszerekkel tudnánk különválasztani egy palimpszeszten a két írásréteget, megállapíthatjuk, hogy a feladat lényegében azonos a koktélparti-problémával. Van ugyan két, egymástól független forrás, ám ezeknek csak valamilyen keverékét ismerjük (egy-egy képen mindkét réteg jelen van). A megoldás kulcsa a stanfordi kutatócsoport által előállított képekben rejlik. A két írásréteg más-más összetételű tintával íródott, így a különböző megvilágítású felvételeken más-más arányban jelennek meg.

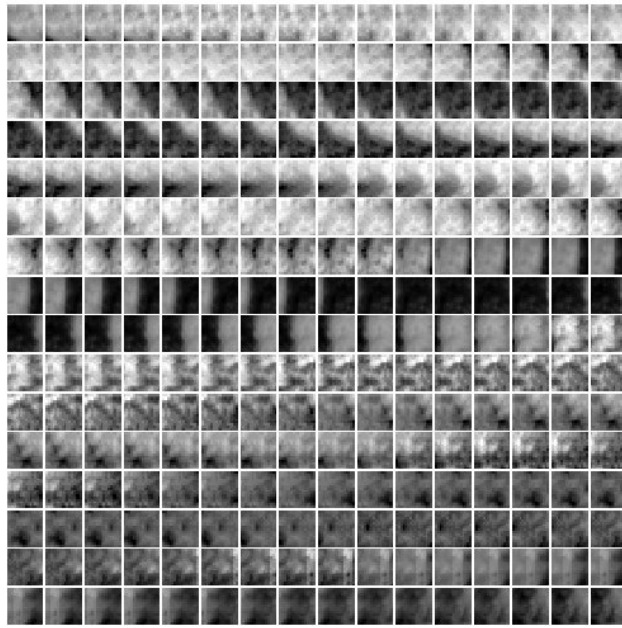
Mivel egyszínű képekről beszélünk, információvesztés nélkül átalakíthatjuk őket szürkeárnyalatúvá. A 3. ábra ugyanazon palimpszeszt egy részletét ábrázolja. Megfigyelhető, hogy a különböző fényű megvilágításnak köszönhetően a két írásréteg más-más arányban jelenik meg a két képen.

3.2. A modell alkalmazása

A koktélparti-problémához hasonlóan, a palimpszeszt esetében is a modell általánosan az 1. képlet szerint írható fel. Megjegyzendő azonban, hogy míg a hangfelvételek esetében az x , illetve s változókat a t idő függvényében indexeltük, a palimpszesztnél – képről lévén szó – a változót mátrixként kezeljük, így sor és oszlop szerint indexeljük. A modell tehát a következőképpen írható át:



3. ábra. Palimpszesztlap részlete



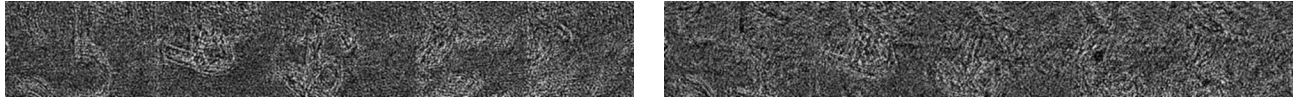
4. ábra. Mintavételezés 16x16-os ablakkal

$$I(x, y) = \sum_{i=1}^m A_i(x, y) s_i, \quad (10)$$

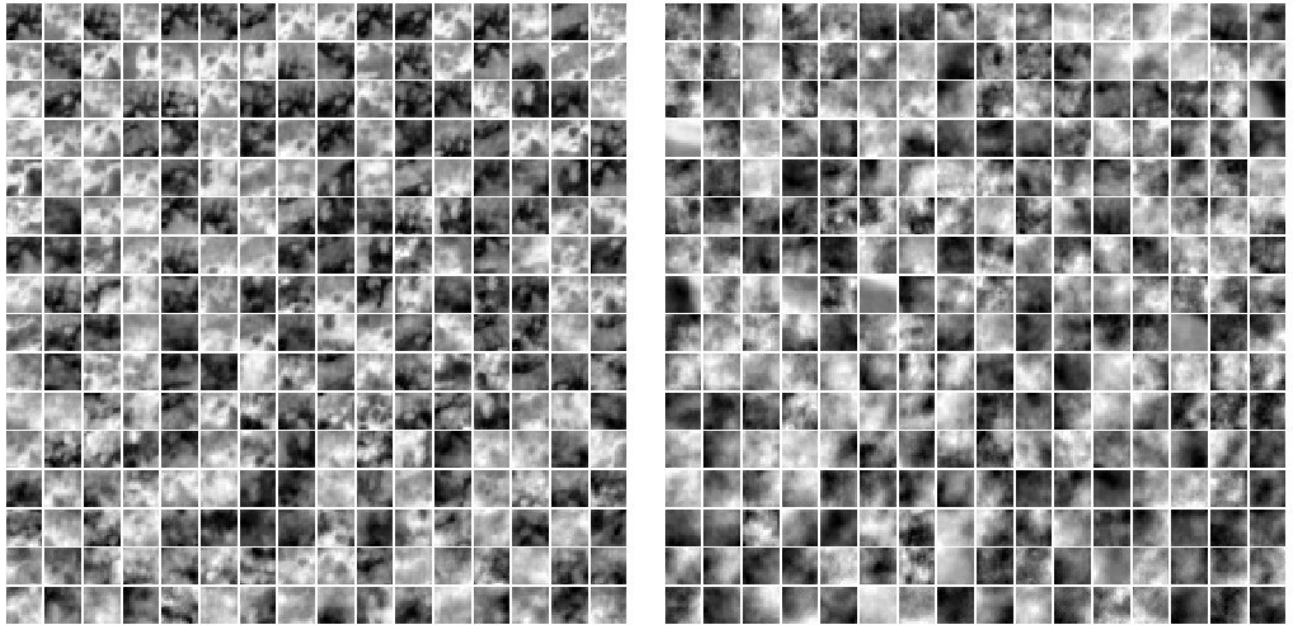
ahol I a megfigyelt képet – vagy annak egy részét – jelöli, s_i a független rétegeket, míg A_i az A keverőmátrix i . oszlopát, m pedig a független komponensek számát. A fenti összefüggés hasonlóan írható fel az összes megfigyelt I képre, mintára

Az I természetesen nem a teljes, 200 MB-os képet jelöli. A mintavételezést 16x16-os “ablakkal” végezzük, amellyel pixelenként végighaladva a képen, kis, négyzet alakú képrészleteket választunk ki. Egy-egy becsléshez átlagosan 50000 mintát használunk. Egy mintavételezés (annak első 16 eleme) például lehet a 4 ábrán látható.

Az így kapott mintákat a hangfelvételekhez hasonlóan oszlopvektorként írjuk fel, tehát az $n \times n$ -



5. ábra. A fastica algoritmus által elkülönített forrásokból kettő



6. ábra. az A mátrix felső "sora", képként ábrázolva (egy kis ábra a mátrix egy oszlopának felel meg)

es mátrixokat $n^2 \times 1$ -es vektorokká alakítjuk. Az ugyanarról a palimpszesztrészletről, ám más-más képről származó vektorokat megfelelően egymás alá "tesszük". A már említett fastica algoritmusnak bemenetként így egy $n^2m \times d$ dimenziójú mátrixot adunk, ahol n a mintavételezéskor használt ablakméret, m a felhasznált képek, illetve keresett független komponensek száma, míg d a minták száma.

A 3. ábrán látott képrészletekre a fastica az 5 ábrán látott eredményeket adta – a szétválasztott komponensmátrix egy-egy sorát megfelelően értelmezve. A rétegek visszaállításához természetesen nem elegendő 1-1 komponens vizsgálata, hanem ezeknek, vagy ezek egy részének valamilyen lineáris kombinációjából állítható elő. A megfelelő komponensek kiválasztása és helyes összevegyítése a további kutatás részét képezi.

Az A oszlopait $n \times n$ -es mátrixokká alakítva (n az ablakméret) és képként értelmezve a 6. ábrán látható mintákat kapjuk.

4. További célkitűzések

Munkám során a következő irányokba szeretnék továbbhaladni:

- a fastica algoritmus kimenetének pontosabb értelmezése;

- a megfelelő komponensek kiválasztása az írásrétegek visszaállításához;
- annak vizsgálata, hogy a bemeneti mátrix dimenziójának növelésével hogyan változik az eredmény;
- egy grafikus felhatalnálói felület létrehozása MATLAB-ban, mely lehetővé teszi (palimpszesz-teket ábrázoló) képek betöltését, a feldolgozni kívánt képrészlet megjelölését, illetve a fastica csomag kimenetének megjelenítését, elmentését.

5. Köszönetnyilvánítás

Ez úton szeretnék köszönetet mondani témavezetőmnek, dr. Csató Lehelnek, akinek szakmai irányítása, felkészítő szemináriumai nélkül e dolgozat nem jöhetett volna létre.

Hivatkozások

- [1] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. on Neural Networks*, 10:626–634, 1999.
- [2] Aapo Hyvärinen, Jarmo Hurri, and Patrik O. Hoyer. *Natural Image Statistics*. Springer, 2008.
- [3] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13, 2000.
- [4] MARC M. VAN HULLE R. MUTIHAC. Comparison of principal component analysis and independent component analysis for blind source separation. *Romanian Reports in Physics*, 56:20–32, 2004.
- [5] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing, Analysis and Machine Vision*. CL-Engineering, 2008.