Benzer ve Aynı Dosyalar

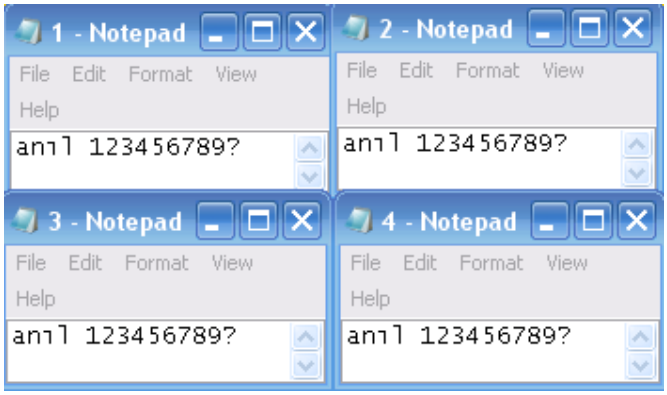
"Herkes tek ve biriciktir, diğer herkes gibi" bu değilse de buna oldukça benzer bir laftı ve bu yazı için güzel bir giriş cümlesiydi, ben de gerekeni yaptım. Bilgisayarımızda aynı dosyaların "isteğimiz dışında" bulunmasından (edebi versiyonu için "at koşturmasından") rahatsızlık duyarız (duyuyorsunuz değil mi). Bu dosyalar yazılar, resimler, sesler ve videolar olabilir.

Aynı dosyaların canımızı sıkma derecesi bizim için anlamlı olup olmadığı ile yakından ilgilidir. Örneğin "C:\Program Files\Common Files\Microsoft Shared\OFFICE12\Office Setup Controller\OSETUP.DLL" dosyasının "C:\MSOCache\All Users\{90120000-0030-0000-0000-0000000FF1CE}-C\osetup.dll" veya herhangi başka bir yerdeki "osetup.dll" ile aynı olması bizi pek bir ilgilendirmez. Anlamlı dosyalarımız içinde ise aynı dosyalara asla katlanamayız, arşivciliğe aykırıdır çünkü. Aynı dosyaları bulmak için yardımcı olan birçok program olmasına rağmen, işimize yarayacak "benzer dosyaları bulucu" programlar pek yoktur. Kısıtlı sayıdaki programlar ise yalnızca tek bir alana yoğunlaşmaktadır.

Aynılık ve benzerlik kavramlarını örnekler üzerinde inceleyelim ilk önce:

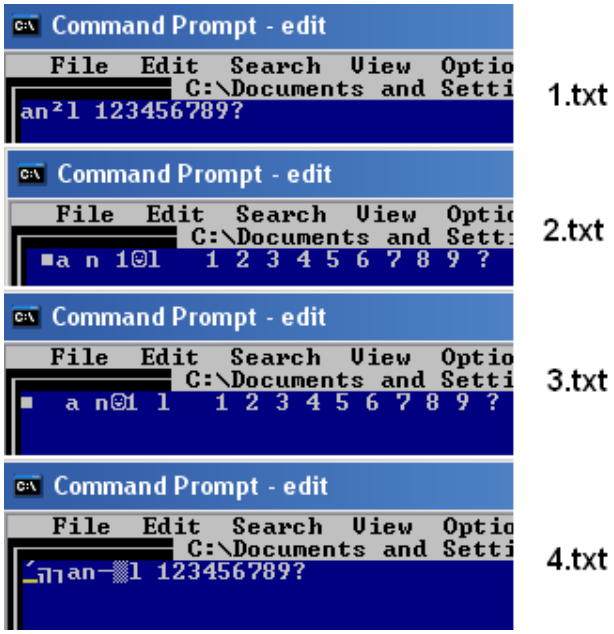
Resim 1. 1 kb lık dört yazı dosyası

Resim 1 deki dosyalar aynı olabilir, çünkü boyutları aynı. Ama bir dosyayı sadece dış görünüşüyle tanıyamayız, içindekileri de görerek onu gerçekten tanırız.



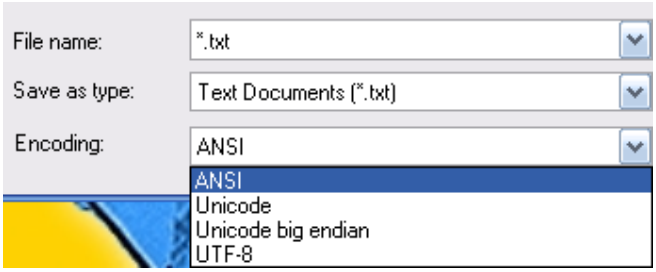
Resim 2. Aynı içerikli dört yazı dosyası

Resim 2 deki dosyalar için artık "aynı olabilir" şüpheciliğinden, "bu dosyalar aynıdır" kesinliğine geçebilir miyiz? Çoğunlukla dikkatli bakmayız ve yanlış fikirlerle gereksiz zaman harcarız. Bilgisayarda hiçbir "ama" kalmayana kadar emin olabiliriz, denemeye devam edelim.



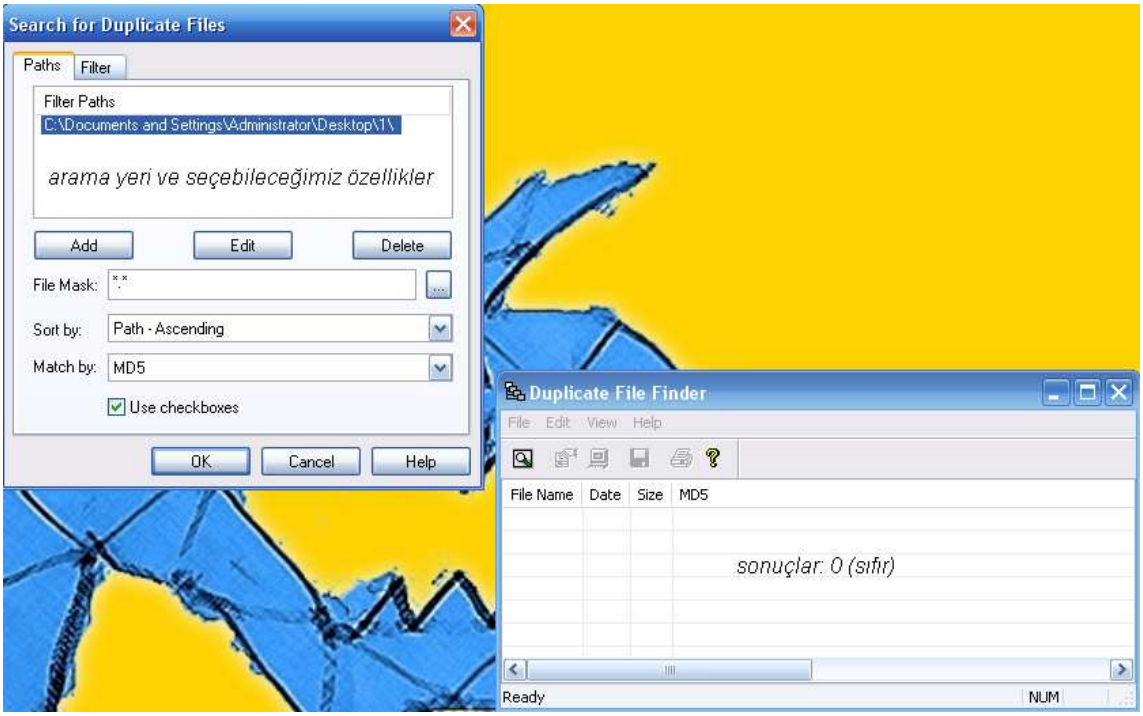
Resim 3. Daha içeriden bir kesit

Dışı bir olsa bile içi farklı gördüğünüz gibi. Bir yazı dosyasını bile farklı kod standartlarıyla kaydettiğimizde yalnızca kendine benzer olduğu için, aynı veya benzer içerikli dosyaların bulunması normal programlama teknikleriyle (veya programlarla) oldukça zordur.



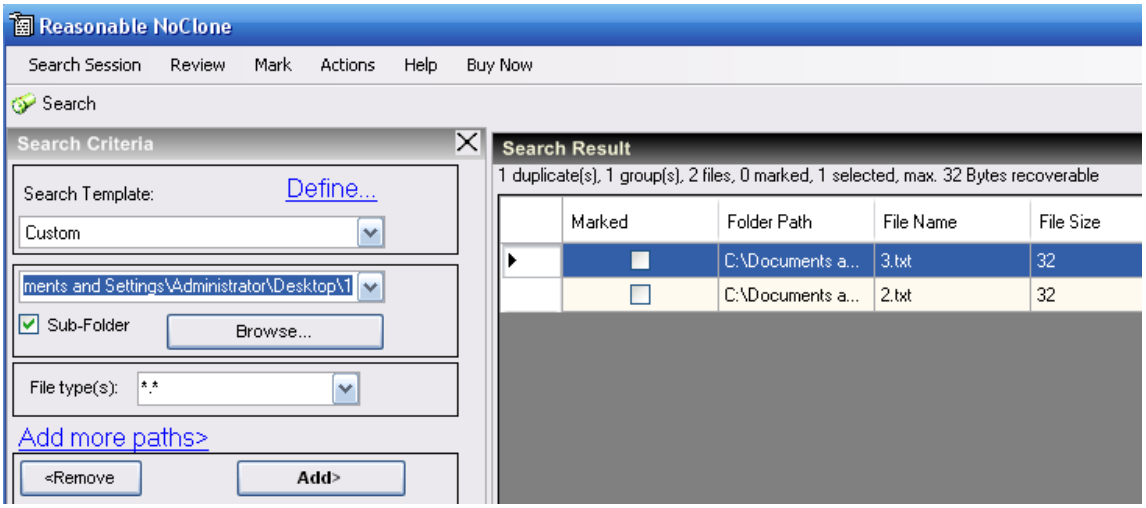
Resim 4. İşin sihri

Örneğin “Dubicate File Finder” isimli program bu dört tane dosyanın aynı olmadığı üzerine yemin ediyor, ya ona inanacaksınız ve mutlu bir şekilde yaşamınıza devam edeceksiniz ya da beni takibe devam edin.



Resim 5. Birebir aynı dosyaları bulan program

Aynı dosyalar üzerinde “hain” deneylerimi şiddetle sürdürdüm. Bu sefer “Reasonable NoClone” isimli programı kullandım. Kullanıcıya daha fazla seçenek sunduğu için (yalnızca binary karşılaştırma yerine, aynı boyutlu dosyaları da bulma özeliği gib) “2.txt” ve “3.txt” nin aynı olduğunu bulabildi.



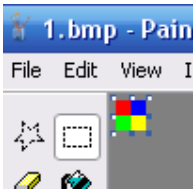
Resim 6. Boyutları aynı dosyaları da bulabilen program

İlk önce kısa bir mazeret yazısıyla giriş yapayım bu paragrafa; internet erişimim kısıtlı imkanlar dâhilinde olduğundan inceleme yapabildiğim programlar ve veriler küçük bir kümeyi kapsar durumda kaldı ama yazının ileriki sürümlerinde (yalana bak!) araştırmayı daha da genişletebilirim. Çünkü eminim ki bir yerlerde bir pdf ile doc un içeriğini karşılaştırabileceğimiz programlar mevcut, aramaya inanmamız kâfi. İncelediğim programlar:

Reasonable NoClone Dubicate File Finder EF Duplicate MP3 Finder Dup Detector VisiPics

Resim 7. Ağıma düşürdüğüm programlar

İçeriğe göre arama hakkında hazırladığım lafları sunmadan önce birkaç resim daha göstermek istiyorum sizlere konunun daha iyi anlaşılmasına katkısı olacağını düşünerek.



Resim 8. 2x2 lik bir resim

1.bmp nin içeriği:

```
BMF_______6___(_____________________________Ä___Ä_____________ÿ__ÿÿ____ÿÿ____
```

1.gif in içeriği:

```
GIF89a____÷_____€___€_€€___€€_€_€€€€€ÀÀÀÿ___ÿ_ÿÿ___ÿÿ_ÿ_ÿÿÿÿÿ_____________
_____________________________________________________________3__f__™__Ì__
ÿ_3__33_3f_3™_3Ì_3ÿ_f__f3_ff_f™_fÌ_fÿ_™__™3_™f_™™_™Ì_™ÿ_Ì__Ì3_Ìf_Ì™_ÌÌ_Ìÿ_ÿ
__ÿ3_ÿf_ÿ™_ÿÌ_ÿÿ3__3_33_f3_™3_Ì3_ÿ33_33333f33™33Ì33ÿ3f_3f33ff3f™3fÌ3fÿ3™_3™
33™f3™™3™Ì3™ÿ3Ì_3Ì33Ìf3Ì™3ÌÌ3Ìÿ3ÿ_3ÿ33ÿf3ÿ™3ÿÌ3ÿÿf__f_3f_ff_™f_Ìf_ÿf3_f33f3
ff3™f3Ìf3ÿff_ff3fffff™ffÌffÿf™_f™3f™ff™™f™Ìf™ÿfÌ_fÌ3fÌffÌ™fÌÌfÌÿfÿ_fÿ3fÿffÿ
™fÿÌfÿÿ™__™_3™_f™_™™_Ì™_ÿ™3_™33™3f™3™™3Ì™3ÿ™f_™f3™ff™f™™fÌ™fÿ™™_™™3™™f™™™™™™
Ì™™ÿ™Ì_™Ì3™Ìf™Ì™™ÌÌ™Ìÿ™ÿ_™ÿ3™ÿf™ÿ™™ÿÌ™ÿÿÌ__Ì_3Ì_fÌ_™Ì_ÌÌ_ÿÌ3_Ì33Ì3fÌ3™Ì3ÌÌ3
ÿÌf_Ìf3ÌffÌf™ÌfÌÌfÿÌ™_Ì™3Ì™fÌ™™Ì™ÌÌ™ÿÌÌ_ÌÌ3ÌÌfÌÌ™ÌÌÌÌÿÌÿ_Ìÿ3ÌÿfÌÿ™ÌÿÌÌÿÿÿ_
_ÿ_3ÿ_fÿ_™ÿ_Ìÿ_ÿÿ3_ÿ33ÿ3fÿ3™ÿ3Ìÿ3ÿÿf_ÿf3ÿffÿf™ÿfÌÿfÿÿ™_ÿ™3ÿ™fÿ™™ÿ™Ìÿ™ÿÿÌ_ÿÌ
3ÿÌfÿÌ™ÿÌÌÿÌÿÿÿ_ÿÿ3ÿÿfÿÿ™ÿÿÌÿÿÿ!ù_______,__________ _¹µ0¢/ _;
```

1.jpg nin içeriği:

```
ÿøÿà ▯JFIF ▯▯▯ ` ` ÿÛ C ▯▯▯▯▯▯▯▯▯▯                ▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯▯
$.' ",#▯▯(7),01444▯'9=82<.342ÿÛ C▯                        ▯▯▯▯▯▯
▯2!▯!2222222222222222222222222222222222222222222222222222ÿÀ ▯▯ ▯ ▯▯▯"
▯▯▯▯▯▯ÿÄ ▯  ▯▯▯▯▯▯▯▯        ▯▯▯▯▯▯▯▯    ▯▯ÿÄ µ▯ ▯▯▯▯▯▯▯▯▯▯▯  ▯}▯▯▯
▯▯▯▯!1A▯▯Qa▯"q▯2▯‘¡▯#B±Á▯RÑğ$3br,       ▯
▯▯▯▯▯%&'()*456789:CDEFGHIJSTUVWXYZcdefghijstuvwxyzƒ„…†‡ˆ‰Š’“”•–—˜™š¢£¤¥¦
§¨©ª²³´µ¶·¸¹ºÁÂÃÄÅÆÇÈÉÊÒÓÔÕÖ×ØÙÚáâãäåæçèéêñòóôõö÷øùúÿÄ ▯▯ ▯▯▯▯▯▯▯▯▯▯
▯▯▯▯▯▯▯▯▯        ▯▯ÿÄ µ▯ ▯▯▯▯▯▯▯▯▯▯▯▯ ▯▯w ▯▯▯▯▯▯▯!1▯▯AQ▯aq▯"2▯▯▯B‘¡±Á
#3Rğ▯br×ð▯
▯$4á%ñ▯▯▯▯▯&'()*56789:CDEFGHIJSTUVWXYZcdefghijstuvwxyz,ƒ„…†‡ˆ‰Š’“”•–—˜™š¢
£¤¥¦§¨©ª²³´µ¶·¸¹ºÁÂÃÄÅÆÇÈÉÊÒÓÔÕÖ×ØÙÚáâãäåæçèéêòóôõö÷øùúÿÚ ▯▯▯ ▯▯▯▯ ? ô+▯▯
øv]>ÚI4▯-İâvfk8É$▯'N(¢ŠÎ—Á▯D7'ÿÙ
```

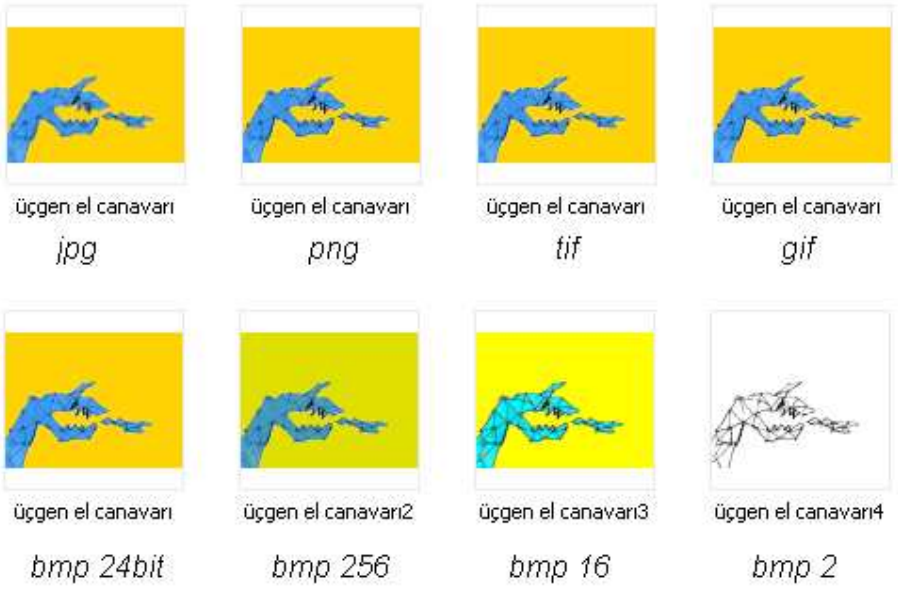
1.png nin içeriği:

```
‰PNG
▯
▯▯▯
IHDR▯▯▯▯▯▯▯▯▯▯
▯▯▯▯ıÔšs▯▯▯▯sRGB▯®Î▯é▯▯▯▯gAMA▯▯±▯
üa▯▯▯▯ cHRM▯▯z&▯▯€„▯▯ú▯▯▯€è▯▯u0▯▯▯ê`▯▯:˜▯▯▯pœ°Q<▯▯▯▯IDAT▯Wcø▯▯▯@â?Ãÿÿ
▯▯ï▯üsÖî3▯▯▯▯IEND®B`,
```

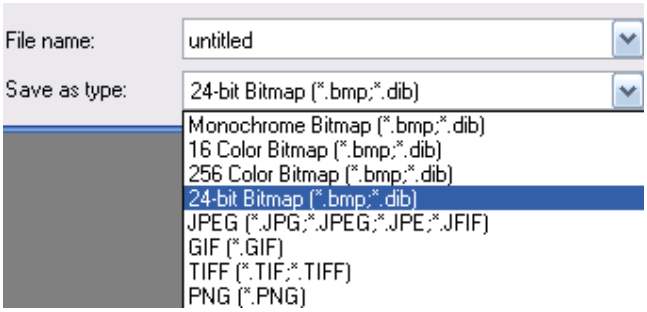
Yukarıda 2x2 lik bir resim dosyası var; birinci satırı kırmızı ve mavi, ikinci satırı ise yeşil, sarı. Anlamsız karakter toplulukları ise bu resmin farklı formatlardaki “kod” hali. Gördüğünüz gibi dosyalar aynı ama farklı yine. Bu tipteki dosyaları bulmak için “içeriğe göre arama” (ben uydurdum literatürdeki ismini bilmiyorum) yöntemlerini kullanmalıyız. Tabiî ki hepimiz doğuştan programcı olmadığımız için kendi algoritmamızı geliştirene kadar hazır programlardan faydalanacağız.

Resim 9. Bir resmin farklı çözünürlükteki kopyaları

Örneğin bir resmin farklı çözünürlüğe sahip kopyalarını iga (içeriği göre arama) yöntemleriyle kolayca bulabiliriz. (Dup Detector ve VisiPics %100 başarı gösterdi bu testte) (Programcı arkadaşlar burada "nasıllar hakkında" biraz açıklama yaparlarsa biz "meraklı tüketicileri" sevindirirler. Ben kendi tahminlerimi yazayım, dosyalar hızlı bir formülle ortak bir çözünürlüğe çevrilir ve karşılaştırma yapılır veya resimlerin örneğin yukarıdan %20 içeri, soldan %25 gibi bölgelerindeki renk kümelerine bakarak karşılaştırma yapılabilir,…) Bir insan için "çocuk oyuncağı" olan bu işi yapabilen yalnızca iki programın olması bence ne anlam ifade ediyor kısmı sonuç kısmında yer alacak.

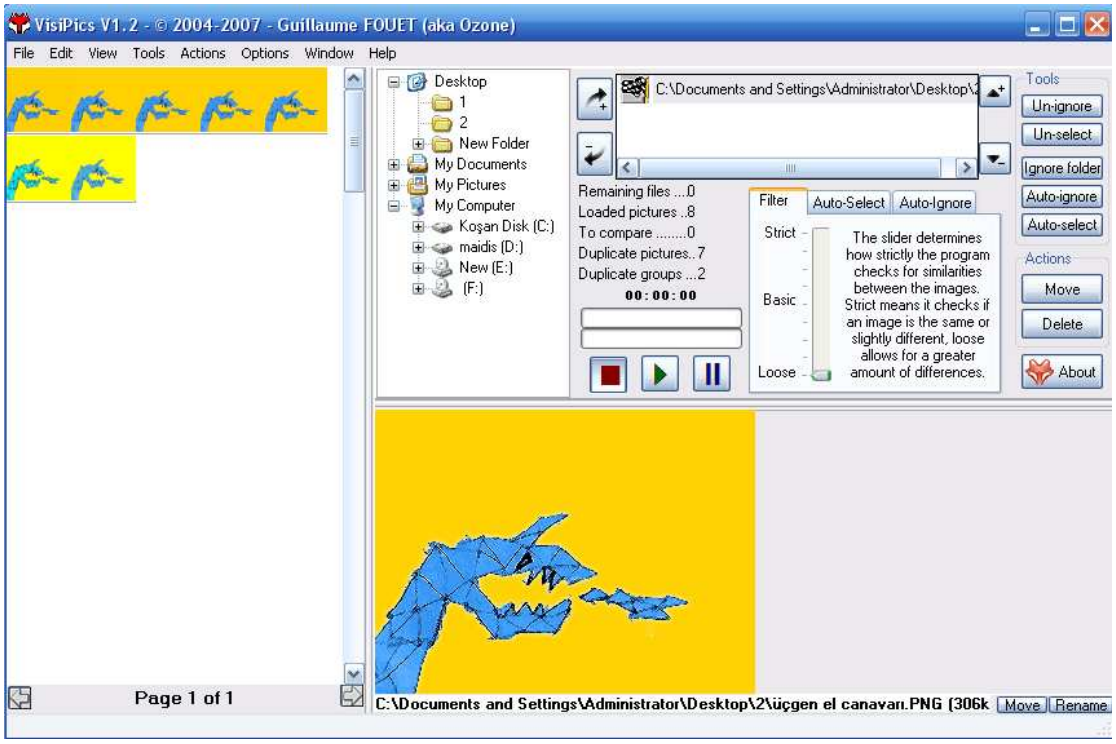


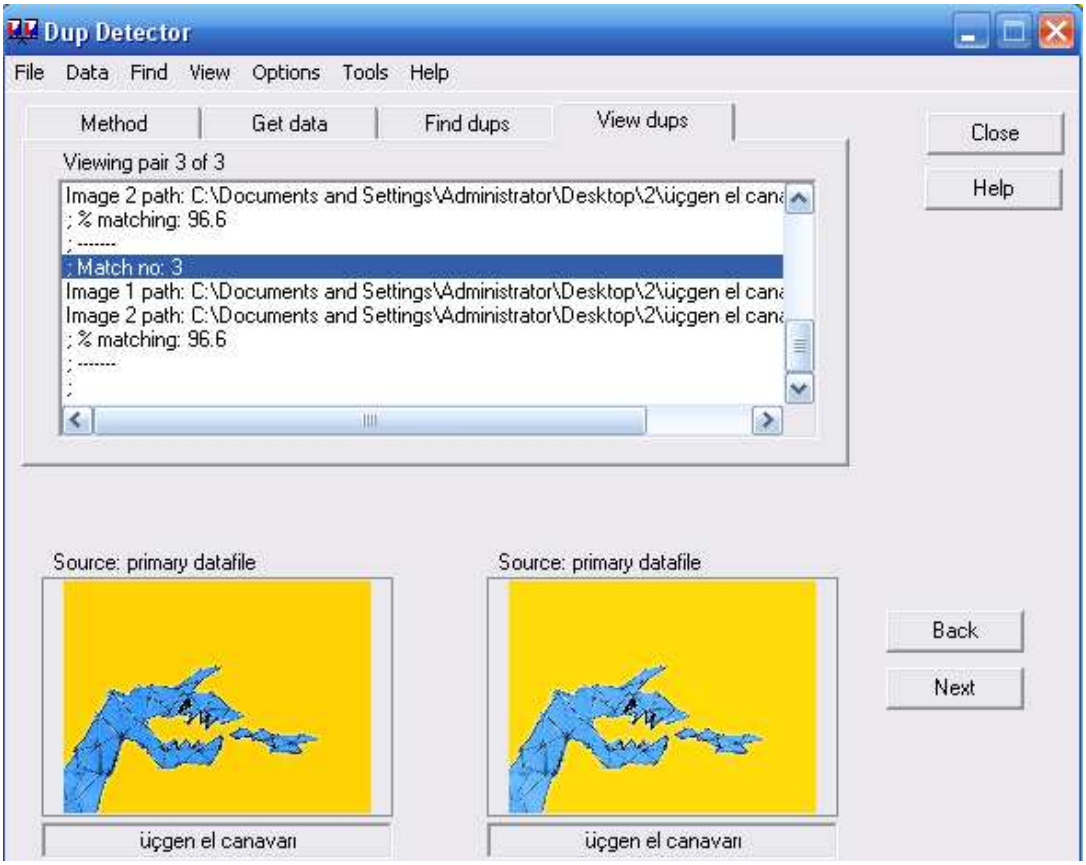
Resim 10. Aynı dosya farklı formatlarda



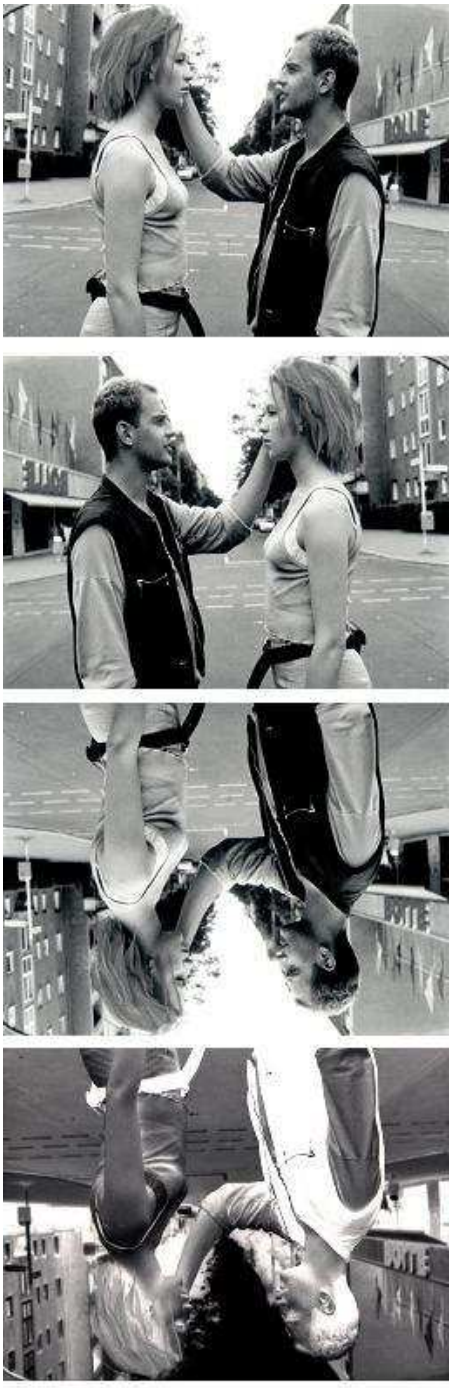
Resim 11. Bazı görüntü formatları

Sonraki testte bir dosyayı farklı formatlarda kaydettim ve yaptığım test sonuçları aşağıdaki gibidir. (Dup Detector bu sefer tatmin edici sonuçlar veremedi, VisiPics ise görevini başarıyla tamamlamanın mutluluğunu duyumsayamadıysa da benden bir aferin aldı)





Resim 12. Test sonuçları



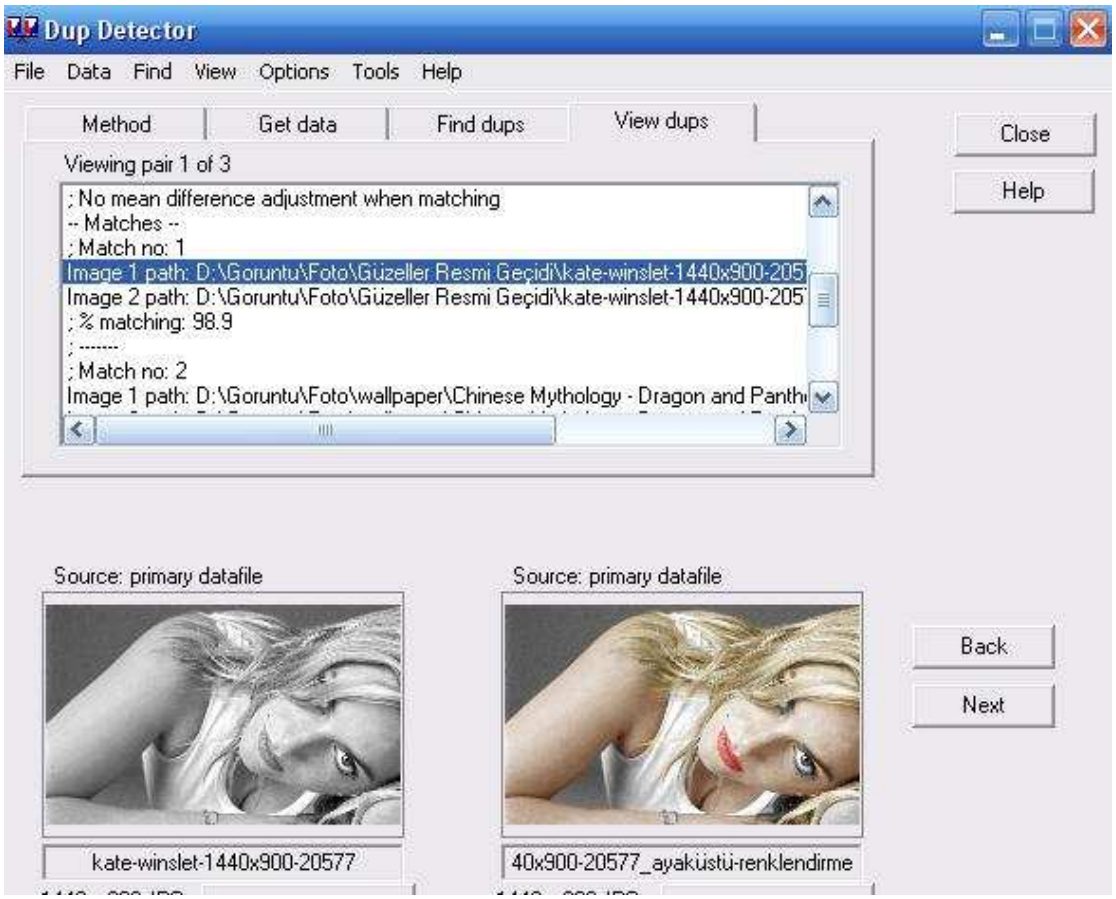
fotoğrafın orjinal ismi mb-fp-bw-1.jpg dir sahibini bilmiyorum

sonradan gelen edit: sanırım run rola run

Resim 13. Bazı bubi tuzakları

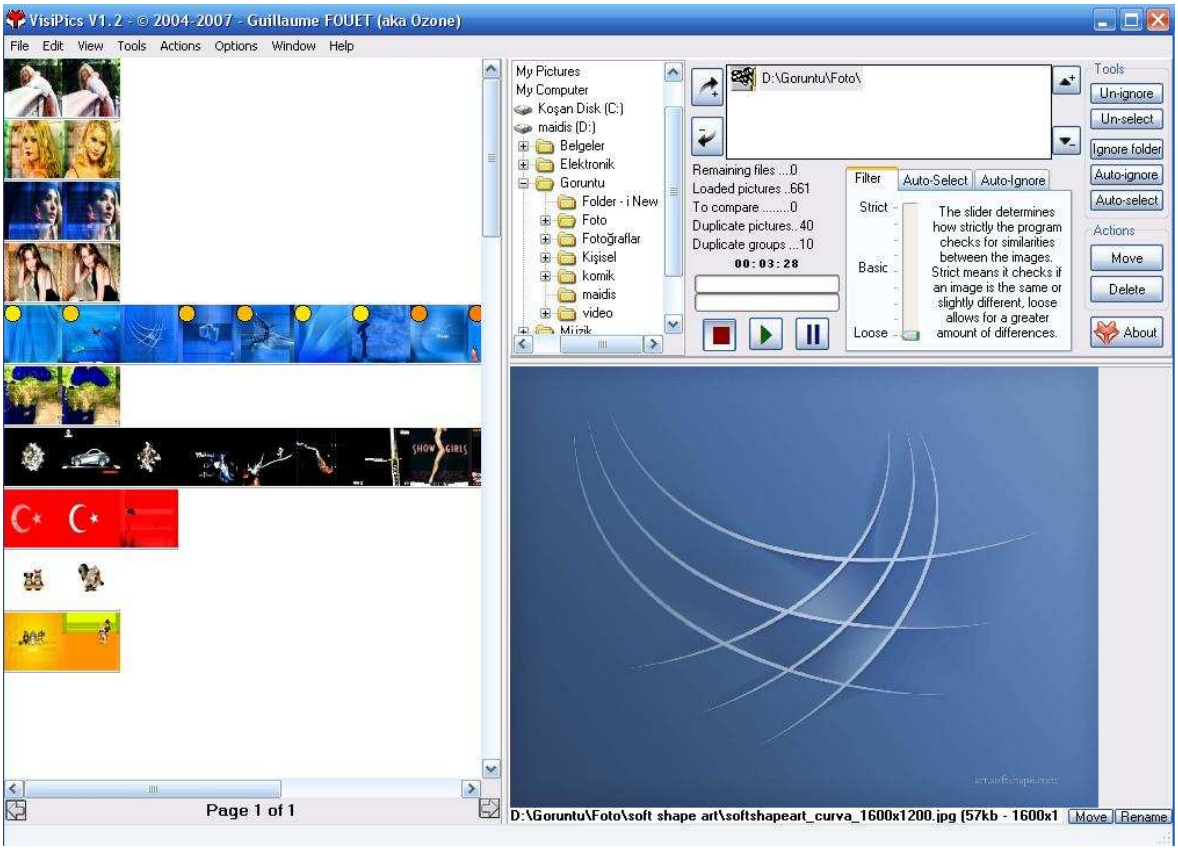
Daha da sonra bir resmin yönünü değiştirerek, aynalayarak ve renklerinin tersini alarak çeşitli denemeler yaptım, Dup Detector bu sefer sınıfı zar zor geçerken, VisiPics hiçbir varlık gösteremedi.

Bir sonraki testte ise internetten indirdiğim ve içeriğini genel olarak tanımlayabileceğim “foto” klasöründe arama gerçekleştirdim. Sonuçlar:



Resim 14. Dup Detector ün bulduklarından

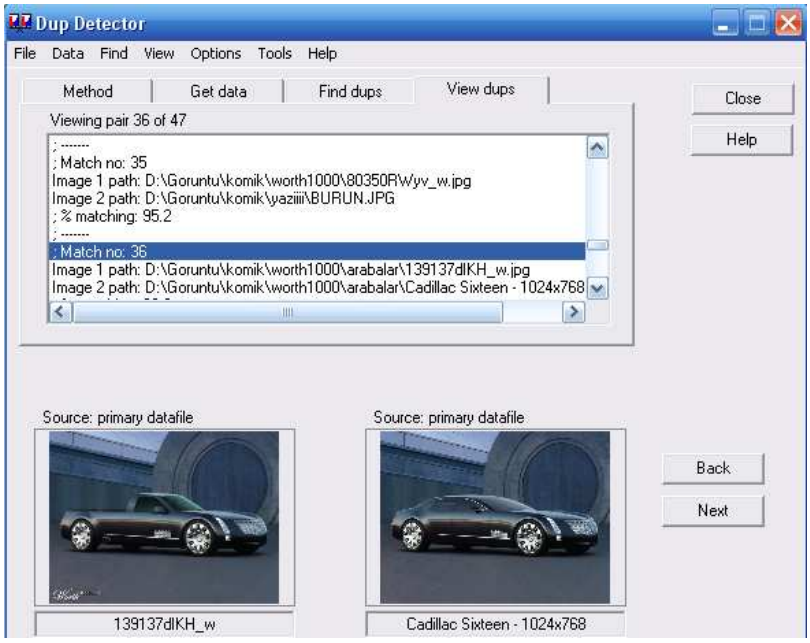
Aynı fotoğrafın renkli ve siyah beyaz kopyasını başarıyla bulan Dup Detector, genelde moda çekimlerinde karşımıza çıkan seri fotoğraflarda oldukça silik kaldı



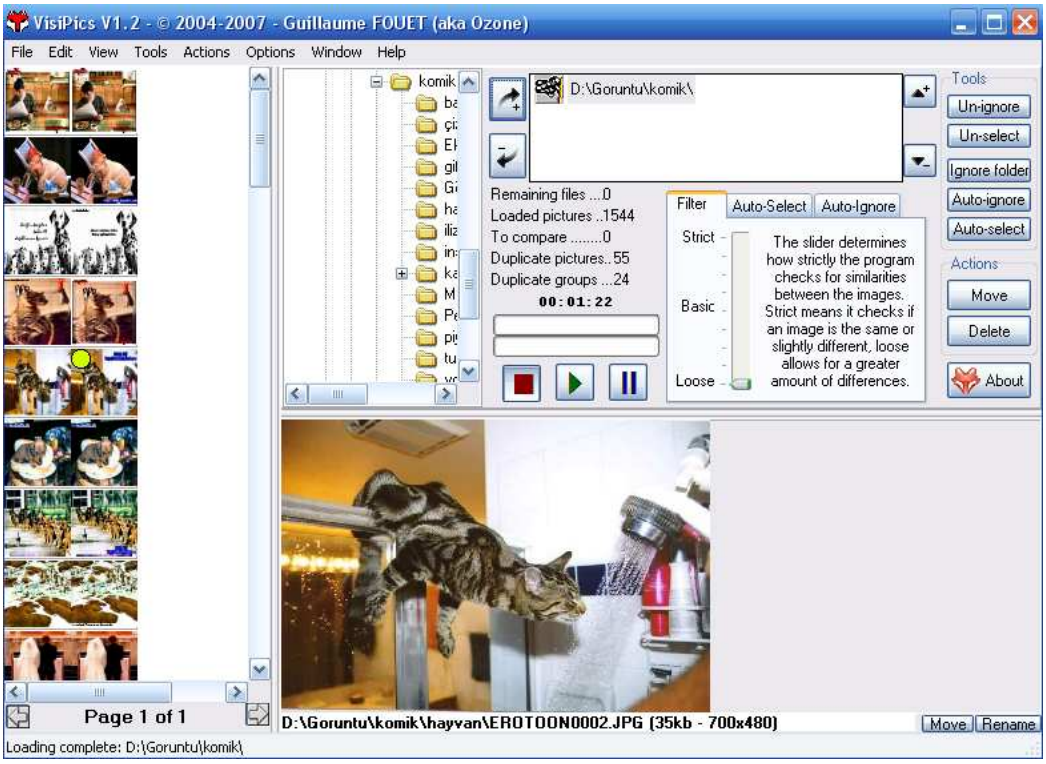
Resim 15. VisiPics in bulduklarından

Bu klasörde harikalar yaratan VisiPics, bulması gereken tüm dosyaları listeleyerek gönüllerdeki yerini biraz daha sağlamlaştırdı.

Birçok bilgisayar kullanıcısının bilgisayarının bir yerlerinde konumlandırdığı ve içinde karikatür, montaj fotoğrafları gibi eserlerin yer aldığı “komik” ya da benzer bir isimde klasörü vardır. Şimdiki testimiz tam da burada.

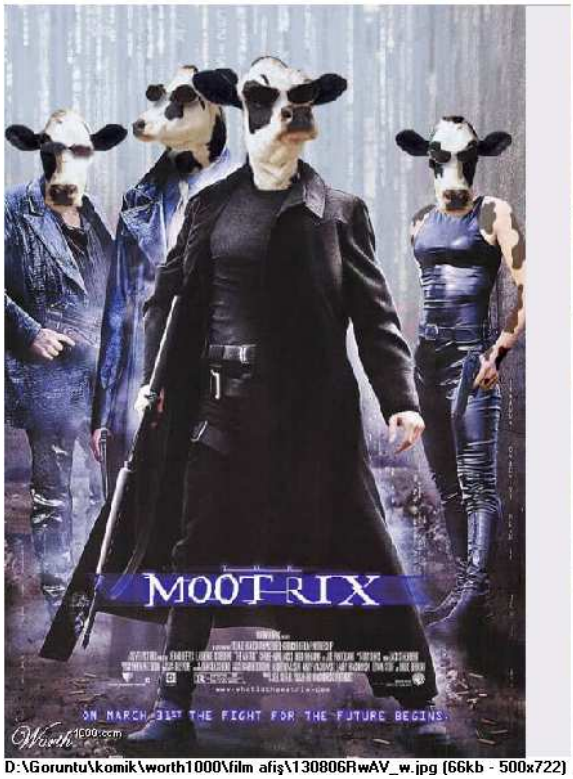


Resim 14. Hanım koş Dup Detector bir şeyler buldu

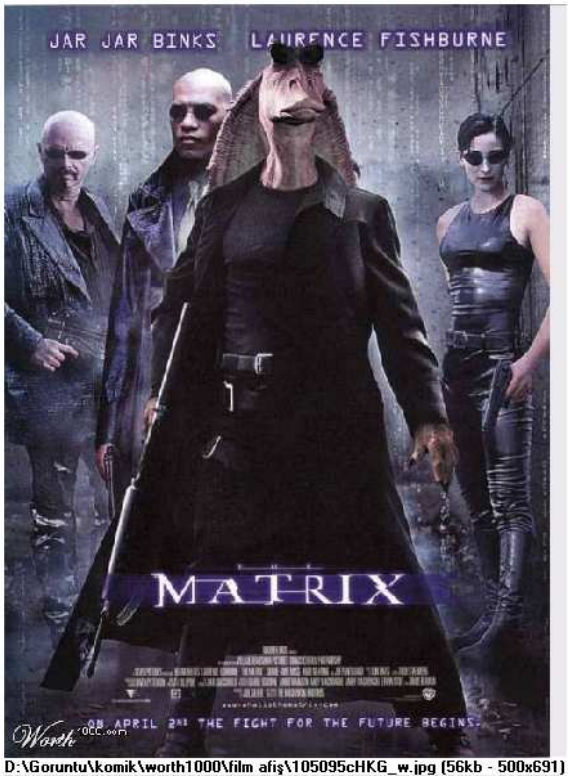


Resim 14. Kedidir kedi

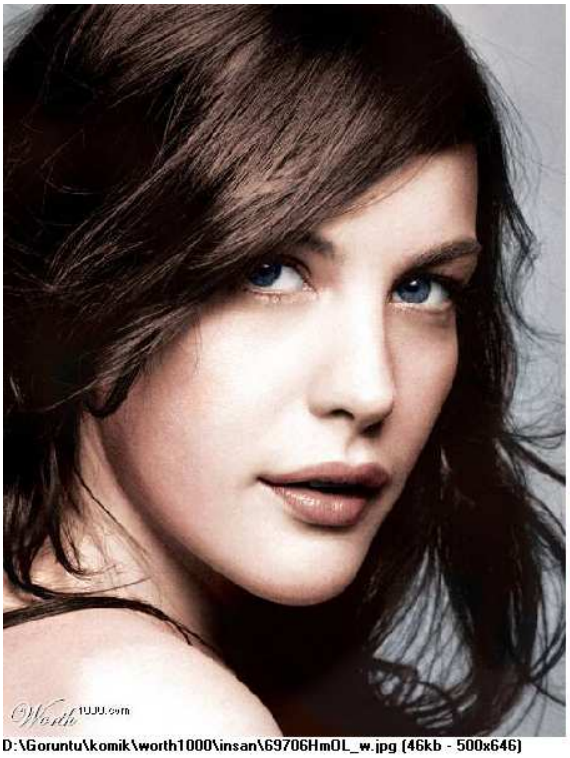
Bu testi açık ara önde olarak VisiPics kazandı. Hiçbir dosyayı ıskalamadı. VisiPics in bulduklarına daha yakından bir bakış için:



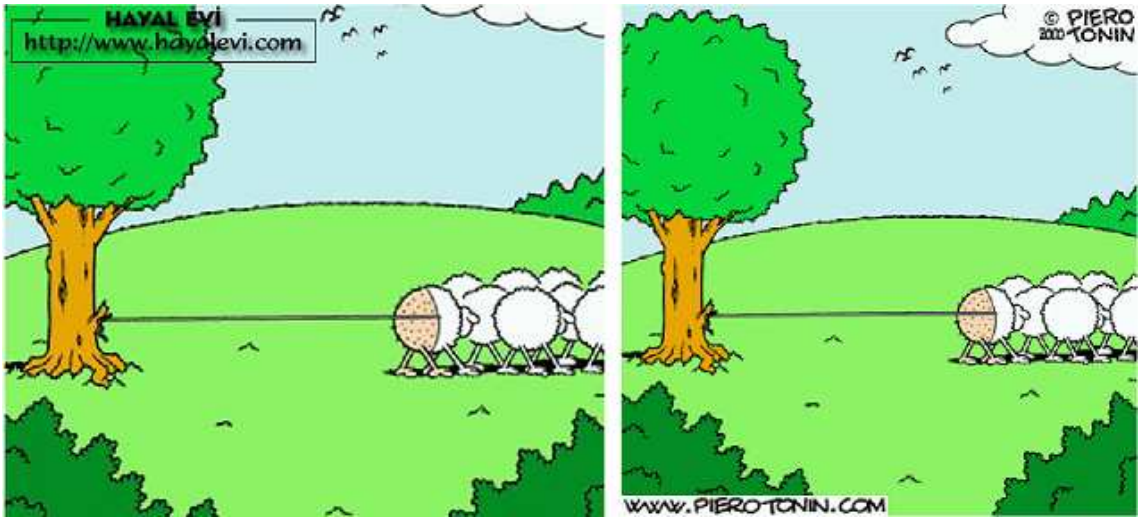
D:\Goruntu\komik\worth1000\film afiş\130806RwAV_w.jpg (66kb - 500x722)



D:\Goruntu\komik\worth1000\film afiş\105095cHKG_w.jpg (56kb - 500x691)


D:\Goruntu\komik\worth1000\insan\69706HmOL_w.jpg (46kb - 500x646)

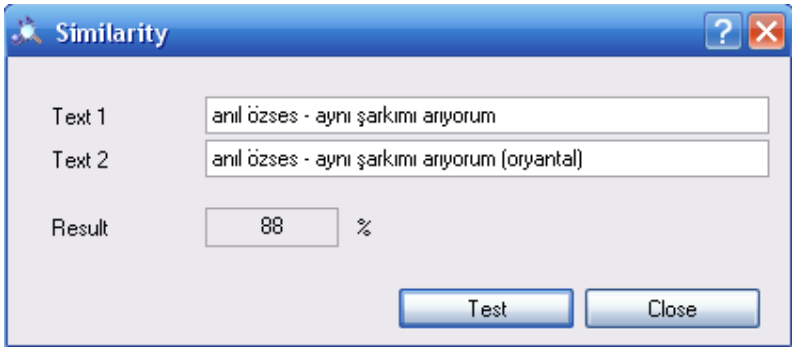

D:\Goruntu\komik\worth1000\insan\69699nsUR_w.jpg (44kb - 500x644)



Resim 15. Tıpkısının aynısının biraz farklısı

Tek karelik görüntüler için tatminkâr aramalar yapılabilmektedir. Bu teknolojilerin biraz daha geliştirilmesiyle hareketli görüntüler için de kolayca uyarlanabilir (Örneğin hareketli bir görüntünün kendi içinde yeterli farklılığa sahip (tamamen aynı renk değil, vs…) bir karenin alınarak diğerleriyle karşılaştırılması zaman alıcı gibi görünse de ilk akla gelen yöntem oldu).

Ses dosyaları için ise durum bu kadar parlak değil ya da ben yanlış yerlerde aradım hep. Demek istediğim ses analizi yaparak benzer dosyaları bulan program yok (yeteri kadar araştırma yapmadığım için bu lafımı yemeye seve seve razıyım). Klasik birebir arama yapan programlar haricinde şarkıcı ismi, parça ismi gibi anahtar ifadelere bakarak arama yapan programlar mevcut sadece.



Resim 16. Hangi yöntem

Aynı ses dosyalarını bulan arama programlarındaki “şimdilik” en işe yarar yöntem şarkıcı ismi ve parça ismi benzer olan (bkz: resim 16) dosyaları eşleştirmektir. Aynı dosyaları bulması ise sadece arşivinizdeki mp3 bilgilerinin doğru girilmiş olmasına ve şansa bağlıdır.

Sonuç kısımları her zaman zorlandığım bölümler olmuştur. Kısa keseceğim. Çoğu kişinin şu anda ihtiyaç bile duymadığı bu yazılımlar ileride çok önemli bir konumda olacaktırlar. Bilgi beraberinde bilgi kirliliğiyle birlikte hızla artmaktadır. Bir makaleyi veya şiiri okuyarak birbirinden esinlenip esinlenilmediğinin tespiti, fotoğraf ve resimleri inceleyerek taklit edilip edilmediği gibi şeyler ve daha birçoğunun çıkış noktası olacağını düşünüyorum bu konunun.

Maidis